Preprints.org

Article

Not peer-reviewed version

A Multistage Binomial Model with Measurement Errors: Application to Protein Viability Prediction

Chénangnon Frédéric Tovissodé *, Yesol Sapozhnikov, Craig Miller

Posted Date: 30 September 2025

doi: 10.20944/preprints202509.2376.v1

Keywords: multistage process; binomial regression; Berkson measurement error; Jeffreys invariant prior; mutation viability

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

A Multistage Binomial Model with Measurement Errors: Application to Protein Viability Prediction

Chénangnon Frédéric Tovissodé [1,*], Yesol Sapozhnikov [2] and Craig Miller [1,3]

[1] Institute for Modeling Collaboration and Innovation, University of Idaho, Moscow, ID 83844, USA
[2] Program in Bioinformatics and Computational Biology, University of Idaho, Moscow, ID, 83844, USA
[3] Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA
* Correspondence: ctovissode@uidaho.edu

Abstract

Many systems exhibit multistage mechanisms where failure of any single component leads to overall failure. Standard logistic regression, with its additive log-odds structure, is not well suited for such data-generating processes. We propose the Multistage Binomial (MSB) model, an extension of the binomial generalized linear model in which component effects combine multiplicatively on the success-probability scale. The MSB model naturally accounts for unobserved components by allowing the success probability to asymptote below one. It additionally incorporates measurement variability in explanatory variables through a Berkson-type error framework. We establish conditions for identifiability, develop a penalized maximum likelihood estimation procedure, and a non-standard likelihood ratio test for unit asymptote. Using synthetic data on mutated protein viability, we show that the likelihood ratio test is conservative, and only rejects the unit asymptote assumption when it is strongly supported by data. We also demonstrate that the MSB model provides more accurate inference and prediction than traditional logistic regression in multistage settings. Our simulations further show that larger sample sizes are required with increased proportion of unobserved components. We provide a chart for sample size determination under an MSB design for real data analysis given the desired accuracy and the distribution of predictors.

Keywords: multistage process; binomial regression; Berkson measurement error; Jeffreys invariant prior; mutation viability

1. Introduction

Binomial data are routine in quantitative sciences, naturally arising for instance in biology (e.g. presence or absence of a particular disease, a mutation leading to a viable or an unviable protein, survive or not survive the great migration in the Serengeti), economics (e.g. purchase or not purchase a product), material science (chiral or achiral compound, presence or absence of defects), and behavioral science (e.g. voting, school dropout, divorce). Binomial regression is the modern approach to analyzing such dichotomous outcomes, and most researchers turn to the logistic model [1–3].

In the Generalized Linear Model (GLM) framework [4], the logistic regression model for a binary outcome Y assumes a log-linear odds ratio for the success event $Y = 1$ versus the failure $Y = 0$, i.e. the expectation μ (also the success probability) of Y satisfies

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_1 X_p \tag{1}$$

where X_j ($j = 1, \cdots, p$) are explanatory variables for Y, and β_j ($j = 0, 1, \cdots, p$) are regression coefficients. Stated, otherwise, the effects of the predictors of the odds ratio $\frac{\mu}{1-\mu}$ of the event $Y = 1$ are multiplicative. The logistic assumption (1) is mathematically convenient, allowing an unrestricted parameterization of the binomial GLM: the coefficients β_j (which measure changes in the log-odds ratio) take values on the entire real line.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Despite the popularity of the standard logistic regression (multiplicative odds) for binary outcomes, the underlying assumption of this model does not cover phenomena resulting from many independent or successive discrete events. Indeed, complex systems often progress through a series of rate-limiting steps which can be subject to modifications by environmental exposures [5]. There can thus be a large number of routes to the failure of such systems, but only the final state is often observable. Indeed, many physical processes can be described by a number of specific events occurring randomly or in some predetermined order. In other words, the occurrence of a particular event often depends on the prior occurrence of many other (partially) related events.

Examples of such events are common along a manufacturing chain: the absence of defects on an end-line product obtained from assembling q different pieces depends on the occurrence of defects in each assembly line producing a piece. If the lines are independent, then the probability μ of a defect-free final item is the product of the no-defect probabilities μ_j in the q lines ($j = 1, \cdots, q$):

$$\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_q. \tag{2}$$

Of course, the design of products is generally optimized so as to make errors additive and the effect of a single defect as insignificant as possible. However, if only the presence or absence of defects is of interest, then the multiplicative absolute risk model (2) reflects the process generating the binary outcome.

Processes with multiplicative risks are in fact common in biology, since many biological processes involve discrete events, at least at some levels of description and observation [6,7]. Consider, for instance, a mutated protein resulting from a single amino acid substitution, with functional viability depending upon two factors: the protein's ability to fold into a native conformation (related to folding stability), and its ability to bind other subunits or ligands (related to binding stability). If such a protein does not fold properly (e.g. because of a highly destabilizing mutation), it is likely not going to bind (very low probability) since most proteins only bind their targets when folded [8]. The viability of the mutated protein is thus proportional to the probabilities of folding (μ_1) and binding (μ_2): $\mu \propto \mu_1 \times \mu_2$ so that any predictor value leading to a small probability begets an overall prediction near zero (i.e. a mutation that renders a protein able to fold, but not bind, or vice versa, should be inviable). In this specific case, the multiplicative odds model (1) appears particularly inappropriate. Indeed, for $p = 2$ with X_1 and X_2 representing folding and binding stabilities, respectively, if we fix folding stability to a value x_1, then it is always possible to find some value x_2 of binding stability so that the viability gets arbitrarily close to one (including an interaction term $\beta_{12}X_1X_2$ into (1) does not help because this is simply equivalent to adding $\beta_{12}x_1$ to β_2). This contradicts the fact that when folding stability is disrupted, the functional viability of the mutated protein cannot get arbitrarily close to one, regardless of the binding stability.

More generally, a multiplicative risk model arises naturally for any all-or-none type event. For instance, the presence or absence of a disease may result from the total absence of one or more protective events (e.g. diet deficiencies in required dietary constituents leading to a resultant disease) [9,10]. Conversely the presence or absence of a particular disease may result from one or more adverse events (e.g. encounters with carcinogenic agents leading to a specific cancer). Multiplicative risks are indeed widely used, for instance, in multi-stage cancer models [11–17], and for incidence or hazard rate models in survival analysis [18–20]. In social epidemiology, many studies reported evidence for multiplicative relations between the effects of risk factors and disease incidence or mortality [21–25].

However, to the best of our knowledge, there is no general binary data model with multiplicative absolute risks. To overcome interpretation difficulties related to the non-collapsibility of odds ratio [26–30], some works considered log-binomial models where, as in a Poisson GLM, covariates have multiplicative effects on the relative risk [31–34]. In addition, the Zero-Inflated Bernoulli (ZIB) model, a special case of the zero-inflated binomial model [35], is related to the multiplicative risk model (2); as we discuss in Section 2, the ZIB model can be described as a special case of our proposal. In spite of some recent developments in relative risk modeling where the probability of the target

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

event is modeled as a ratio to the probability under a reference condition [36,37], these models not only lack the mathematical convenience of the logistic model (sample dependent constraints required on the parameter space to ensure that predicted probabilities are restricted to $[0,1]$, or restricted ranges of covariates), but also have their own set of interpretation issues [38–41].

We propose in this work an extension of the Traditional Binomial (TB) model to handle binomial data resulting from data generation processes with known multiplicative risk components. The proposed model assumes that in addition to an observed binomial outcome, we are aware of some unmeasurable prior events that determine the final outcome (with paths to success or failure involving sequential, or non-sequential steps, or a combination of both), and each prior event is known to be related to a subset of the observable explanatory variables. The proposal results from combining equation (2) with a standard binomial model inverse link function, i.e. each factor μ_j is given by $\mu_j = h(\eta_j)$ with $h : \mathbb{R} \to [0,1]$ and $\eta_j \in \mathbb{R}$ a linear predictor obtained from explanatory variables related to μ_j. Most popular binomial model link functions include the logit link function (1) with inverse the cumulative distribution function (cdf) of the logistic distribution,

$$h(\eta_j) = \left[1 + \exp(-\eta_j)\right]^{-1}, \tag{3}$$

and the probit link with $h(\eta_j) = \Phi(\eta_j)$ where Φ is the cdf of the standard normal distribution. The resulting model is termed Multistage Binomial (MSB) model. It is worthwhile emphasizing that the MSB model (with multiple stages) is not a generic model for binomial data. Its use should be prescribed by some theoretical or empirical knowledge of the phenomenon being modeled. In case of only a (theoretical) suspicion of multiple stages, the MSB model should be tested against the TB model.

In addition to providing a general framework for modeling multiplicative risks in binomial data, nesting the TB and ZIB models, our proposed MSB model allows to control for two important sources of issues in statistical analysis: omitted variables and measurement errors in predictors. Indeed, researchers can seldom be aware of, let alone measure, all of the variables that determine a target outcome, and the impact of left out or omitted variables can be dramatic for interpretations in TB models [42–50]. Our formulation of the MSB model allows to control for omitted variables in some variable omission scenarios, i.e., when the explanatory variables which determine some factors μ_j in (2) are omitted. Moreover, observed predictors are often error-prone. To account for errors in observed explanatory variables, the proposed MSB model also allows for heteroscedastic additive Berkson measurement errors [51] when the variability of each numerical predictor around its observed values is known.

For inference in the MSB model, we developed a Penalized Maximum Likelihood (PML) estimation procedure. The PML method has two main advantages over a simple Maximum Likelihood (ML) approach: reduce *small sample bias* and circumvent estimation issues related to quasi or complete *separation*. On the one hand, although ML estimates in TB regression are asymptotically unbiased to first order, it is well known that the ML estimators are biased to first order in small sample ($n < 200$ [52,53]). This occurs especially when the marginal success probability of the response approaches zero or one [54]. The ML estimators of the MSB model parameters are also expected to be biased to the first order in small samples because this is a general feature of binomial regression models [55].

On the other hand, complete separation occurs when one or more of the explanatory variables can perfectly predict some binary outcomes [56]. Quasi separation occurs when for any values of an explanatory variable (or a linear predictor) x_i less than a threshold, say $x_i < t$, the binary outcome is always $y_i = 0$ (or always $y_i = 1$), but when $x_i \geq t$, y_i might equal zero or one. The occurrence of (quasi) separation is generally related to small sample, sparse data, rare outcome ($E[Y_i]$ is very small), rare exposures (and more generally predictors with very small variability), large number of dichotomous risk factors [57], and highly correlated predictors (multicollinearity) [58], or covariates with strong effects [59].

Since the MSB model can be viewed as a combination of many TB regressions across stages of a complex process, the occurrence of separation becomes more likely not only as the number p of

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

predictors in any stage increases, but also as the number q of stages increases, for not sufficiently large samples. As a consequence, in the MSB model framework, we anticipate an increased likelihood of unrealistically large ML estimates or related large (or undefined) standard errors as p or q increases, and as the sample size n or the maximum success probabilities decrease. Quasi or complete separation implies that maximum likelihood estimates approach infinity or zero [60,61]. In practice, separation generally results in a flat log-likelihood surface (likely with an asymptote) which leads to implausible estimates, and no or very large standard error estimates [62].

In this paper, we investigate some theoretical and numerical properties of MSB models. We first study the identifiability of the proposed model. Then, using simulation experiments, we study finite-sample properties of the PML estimates of the MSB model parameters. Specifically, we *i)* compare, for demonstration purposes, the Two-Stage Binomial (TSB) model to TB and ZIB models considering a simulated biological case study on the viability of a mutated protein; *ii)* assess the ability of PML estimators of MSB to recover the population values of regression parameters and predict event success probabilities in finite samples; *iii)* and evaluate finite-sample properties of a non-standard Likelihood Ratio (LR) test to motivate the choice of the MSB model over alternatives such as the TB or ZIB model.

The organization of the remainder of the paper is as follows. We specify the MSB model, discuss its identifiability and describe the PML inference in Section 2. Then, we present simulation results in Section 3. We discuss our results and finish with concluding remarks in Section 4.

2. The Multistage Binomial Model

The MSB model is built as an extension of the standard binomial GLM [4]. In this section, we specify and describe the main features of the MSB model, starting with the simpler model in the absence of measurement errors in observed covariates, and then introducing measurement errors. We then establish and discuss conditions for the identifiability of the model. Finally, we describe the estimation of model parameters using Jeffreys invariant prior as likelihood penalty, and an LR test to the detect omission of important process stages or predictors.

2.1. Measurement Error-Free Model

Let us consider a design with n units indexed i, each having n_i (known positive integer) independent Bernoulli trials. The binomial outcome Y_i is assumed to be generated through a multistage process with q known steps, each related to a p_j-vector of covariates $\boldsymbol{W}_{ij}$ $(j = 1, 2, \cdots, q)$. The multiplicative risk binomial model is given for the ith outcome Y_i as:

$$(Y_i|\boldsymbol{W}_i = \boldsymbol{w}_i) \overset{ind}{\sim} \mathcal{B}in(n_i, \xi_i), \tag{4a}$$

$$\xi_i = \lambda_i \left[\alpha_i + (1 - \alpha_i) \prod_{j=1}^{q} \xi_{ij}\right], \tag{4b}$$

$$\xi_{ij} = h(\zeta_{ij}), \tag{4c}$$

$$\zeta_{ij} = \boldsymbol{w}_{ij}^{\top} \boldsymbol{\beta}_j, \tag{4d}$$

where $\boldsymbol{W}_i^{\top} = (\boldsymbol{W}_{i1}^{\top}, \cdots, \boldsymbol{W}_{iq}^{\top})$ is the vector of covariates; $\mathcal{B}in(n_i, \xi_i)$ denotes the binomial distribution with n_i trials and success probability ξ_i; $\lambda_i \in (0, 1]$ is the upper limit of the success probability (ξ_i) of each Bernoulli trial given $\boldsymbol{W}_i$ and $\alpha_i \in [0, 1)$ is such that $\alpha_i\lambda_i$ is the lower limit of ξ_i; h is the inverse link function for the binomial outcome (h is assumed to be a continuous cdf, monotone, increasing and twice differentiable); and $\boldsymbol{\beta}_j$ is a p_j-vector of real regression parameters (possibly including an intercept β_{j0}).

Model (4) describes a binomial process where the outcome results from q latent stages where each stage is represented by an independent Bernoulli experiment, and the final outcome corresponds to all experiments being simultaneously successful. As indicated above, the success probability ξ_i is bounded as $\xi_i \in [\alpha_i\lambda_i, \lambda_i]$. These bounds represent theoretical limits that may arise from some unknown stages, or some known stages with no measured explanatory variable. Indeed, setting $n_i = 1$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

without loss of generality, and starting from multiplicative risks of the form (2) with $q+1$ stages ($\xi_i = \prod_{j=0}^{q} \xi_{ij}$) among which the first ($j = 0$) is omitted, the conditional success probability of the outcome given the measured covariates $\boldsymbol{w}_i$ is then

$$\begin{aligned} P(Y_i = 1|\boldsymbol{W}_i = \boldsymbol{w}_i) &= \int_{\mathbb{R}^{p_0}} h\left(\boldsymbol{w}_{i0}^{\top}\boldsymbol{\beta}_0\right) \times \left(\prod_{j=1}^{q} \xi_{ij}\right) f_{\boldsymbol{w}_0}(\boldsymbol{w}_{i0}) d\boldsymbol{w}_{i0} \\ &= \lambda \times \prod_{j=1}^{q} \xi_{ij} \end{aligned} \tag{5}$$

where $\lambda = \int_{\mathbb{R}^{p_0}} h\left(\boldsymbol{w}_{i0}^{\top}\boldsymbol{\beta}_0\right) f_{\boldsymbol{w}_0}(\boldsymbol{w}_{i0}) d\boldsymbol{w}_{i0}$ is the success probability at the omitted stage, averaged over of the distribution of the related p_0-vector $\boldsymbol{W}_{i0}$ of explanatory variables, which has joint density function $f_{\boldsymbol{w}_0}$ (integration will be replaced by summation where $f_{\boldsymbol{w}_0}$ is a probability mass function (pmf), i.e., if $\boldsymbol{W}_{i0}$ is discrete).

Stated in other words, the omission of explanatory variables related to a stage in the multiplicative-risk model (2) leads to (4b) with $\alpha_i = 0$. It is worth noticing that while this rationale applies to unobserved confounders that vary across experimental or observational units, it remains valid when the vector of predictors $\boldsymbol{W}_0$ of a stage is (willfully or not) held constant across all investigated/sampled units: in this case, λ is not an average of success probabilities across the possible values of $\boldsymbol{W}_0$, but rather the conditional success probability of this stage given the constant value $\boldsymbol{w}_0$ of $\boldsymbol{W}_0$: $\lambda = h\left(\boldsymbol{w}_0^{\top}\boldsymbol{\beta}_0\right)$.

This development indicates that λ can arise from one or both of two potential sources: uncontrolled confounders (*marginalization*), or a limit of the data generation mechanism in the specific investigated experimental or observational environment/condition (*conditioning*). In the general formulation (4) of the MSB model, we allow λ to vary across observations to potentially account for *conditioning* when the study is replicated across some discrete experimental/environmental settings: λ_i can capture the maximum success probability for all units i from the same setting. For the sake of symmetry, $\alpha_i \in (0,1)$ is allowed in model (4): since, one can equivalently model the failure of an event to occur instead of its success, the minimum success probability can also be greater than zero. By the same reasoning, this rationale can be extended to the omission and marginalization over many stages, or conditioning on one value of the vector of predictors at each of many stages.

Remark 1. *When model (4) has only one ($q = 1$) or two stages ($q = 2$), we obtain the following two important special cases.*

- *When $q = 1$, $\alpha_i \to 0$, and $\lambda_i \to 1$, the MSB model is reduced to the TB model.*
- *For binary outcomes ($n_i = 1$ for $i = 1, 2, \cdots, n$), when $q = 2$, $\alpha_i \to 0$, and $\lambda_i \to 1$, the MSB model is reduced to the ZIB model.*

  The ZIB model has been proposed by Hall [35] to face situations where a binary outcome contains an excess of zeros as compared to the expected zeros under the TB model. The ZIB model assumes that the binary outcome is the product of two Bernoulli processes; a zero from either process (or both) results in a zero. An example of process generating such data in health sciences was investigated by Diop et al. [63]: a study population consisting of a mixture of susceptible subjects who can experience the binary outcome of interest, and cured subjects (cure fraction) who cannot experience the binary outcome. Complications arise from not knowing the immunity (cured or susceptible) status of each subject, and the ZIB model is used to jointly estimate the cure fraction on the one hand, and the logistic regression for the susceptible subjects on the other hand [64–66]. Using the MSB model framework ($q = 2$, $n_i = 1$, $\alpha_i = 0$, and $\lambda_i = 1$) to describe the ZIB model, the success probability of the binary outcome Y_i is $E[Y_i] = h(\eta_{i1})h(\eta_{i2})$ and the probability that a subject i is cured is $1 - h(\eta_{i1})$. Thus, in this case, the main difference between the MSB model and the ZIB model is that the MSB model allows the minimum probability $\alpha_i\lambda_i$ to be greater than zero, and the maximum success probability λ_i to be less than one.

  When the susceptibility probability $h(\eta_{i1})$ is constant across subjects in the ZIB model, the latter can alternatively be considered as a one stage model ($q = 1$) with $\alpha_i = 0$ and $\lambda = h(\eta_{i1})$. In this case, the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

success probability of the binary outcome Y_i is $E[Y_i] = \lambda h(\eta_{i1})$ and the probability that a subject i is cured is $1 - \lambda$.

In complement to Remark 1, it is worthwhile emphasizing that the MSB model is not a zero inflation model, nor a mixture model for count data. These models only share some special cases with the MSB model when each and every count is binary. For general binomial data, the zero-inflated binomial model [67] is not a special case of the MSB model, even when $q = 2$.

Remark 2. *An ovulation-time model was developed by Garel et al. [68] to investigate how the probability of ovulation and body growth co-evolve over time in moose. The special case of the probit link MSB model when $q = 2$, $\alpha_i \to 0$ and $\lambda_i \to 1$, is closely related to the ovulation-time model. In the MSB model framework, the first stage in their model represents susceptibility to ovulation at a time, and the second stage is the occurrence of ovulation (this also points to a close relation with the ZIB model). However, unlike the fixed-scale link used in proper binomial models, the ovulation-time model employs a covariate-dependent probit link scale to capture spatiotemporal dynamics in the occurrence of ovulation.*

2.2. Multistage Binomial Model Specification

In practice, the covariates $\boldsymbol{W}_i$ in model (4) can be observed with measurement errors. We consider a situation where $\boldsymbol{W}_i$ is error-prone: instead of $\boldsymbol{W}_i$, we only have access to an observable $\boldsymbol{X}_i^\top = (\boldsymbol{X}_{i1}^\top, \cdots, \boldsymbol{X}_{iq}^\top)$ such that $\boldsymbol{X}_{ij}$ satisfies:

$$\boldsymbol{X}_{ij} = \boldsymbol{W}_{ij} + \boldsymbol{e}_{ij} \tag{6}$$

where $\boldsymbol{e}_{ij} = \left(e_{ij1}, \cdots, e_{ijp_j}\right)^\top$ is a measurement error vector independently following an unspecified zero-mean distribution,

$$\boldsymbol{e}_{ij} \overset{ind}{\sim} ME(\boldsymbol{S}_{ij}), \tag{7}$$

which has covariance matrix $\boldsymbol{S}_{ij}$ (a known semi-positive definite matrix), and a probability density function (pdf) $f_{\boldsymbol{e}}(\cdot|\boldsymbol{S}_{ij})$. The measurement error e_{ijk} in (6) is known as an additive Berkson error [51]. This error model is appropriate when a subset of the available explanatory variables are obtained from an estimation process such that the observed $\boldsymbol{X}_{ij}$ can be regarded as an estimate of the truth $\boldsymbol{W}_{ij}$, and $\boldsymbol{S}_{ij}$ is an estimate of the covariance matrix of $\boldsymbol{X}_{ij}$.

For inference purposes here, we are interested in the expected value of the outcome Y_i given an observed covariate vector $\boldsymbol{x}_i$: $E(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = n_i\mu_i$ where n_i is the number of trials, and μ_i is the success probability given $\boldsymbol{X}_i = \boldsymbol{x}_i$. The Multistage Binomial (MSB) model is defined as:

$$(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) \overset{ind}{\sim} \mathcal{B}in(n_i, \mu_i), \tag{8a}$$

$$\text{with} \quad \mu_i = \lambda_i \left[\alpha_i + (1 - \alpha_i) \prod_{j=1}^{q} \mu_{ij}\right] \tag{8b}$$

where λ_i and α_i are as defined in (4), and $\mu_{ij} = \bar{h}\left(\boldsymbol{x}_{ij}|\boldsymbol{\beta}_j, \boldsymbol{S}_{ij}\right)$ is obtained for $j = 1, 2, \cdots, q$ by integrating out the measurement error vector $\boldsymbol{e}_{ij}$ as:

$$\bar{h}(\boldsymbol{x}_{ij}|\boldsymbol{\beta}_j, \boldsymbol{S}_{ij}) = \int_{\mathbb{R}^{p_j}} h\left(\eta_{ij} - \boldsymbol{e}_{ij}^\top \boldsymbol{\beta}_j\right) f_{\boldsymbol{e}}(\boldsymbol{e}_{ij}|\boldsymbol{S}_{ij}) d\boldsymbol{e}_{ij} \tag{9}$$

with $\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j$ (summation replaces integration where $f_{\boldsymbol{e}}$ is a pmf, i.e., if $ME(\boldsymbol{S}_{ij})$ is discrete). From the conditional expectation $E(Y_i|\boldsymbol{X}_i) = n_i\mu_i$, the conditional variance of the binomial outcome Y_i is

$$Var(Y_i|\boldsymbol{X}_i) = n_i\mu_i(1 - \mu_i). \tag{10}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

When there is no measurement error ($\boldsymbol{S}_{ij} = \mathbf{0}$), we have $\mu_{ij} = h(\eta_{ij})$ and $\eta_{ij} = \zeta_{ij}$ from $\boldsymbol{e}_{ij} = \mathbf{0}$, hence $\mu_{ij} = \xi_{ij}$. It follows that the measurement error-free model (4) is a special case of the MSB model (8) when $\boldsymbol{S}_{ij} = \mathbf{0}$ for all stages $j = 1, \cdots, q$; and across all units $i = 1, \cdots, n$.

A design for the general MSB model (8) includes the response vector $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ of n independent outcomes, the associated $n \times p$ (with $p = \sum_{j=1}^{q} p_j$) design matrix $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^\top$, and the array $\mathbf{S}$ of nq known $p_j \times p_j$ covariance matrices $\boldsymbol{S}_{ij}$ of measurement errors. Because the MSB model allows the success probability bounds λ_i and $\alpha_i\lambda_i$ to vary across experimental or observational units, the complete design includes two additional design matrices $\mathbf{U} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n)^\top$ and $\mathbf{V} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n)^\top$ where $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{id})^\top$ is a d-vector and $\boldsymbol{v}_i^\top = (v_{i1}^\top, \cdots, v_{ir}^\top)$ is an r-vector such that $\alpha_i = h(\boldsymbol{u}_i^\top \boldsymbol{\gamma})$ and $\lambda_i = h(\boldsymbol{v}_i^\top \boldsymbol{\delta})$, and $\boldsymbol{\gamma} \in \mathbb{R}^d$ and $\boldsymbol{\delta} \in \mathbb{R}^r$ are parameters that determine the success probability bounds. In particular, each of α_i and λ_i may have a fixed value across units, i.e., we can have $\mathbf{U} = \boldsymbol{J}_n$ (n-vector of all ones) and $\mathbf{V} = \boldsymbol{J}_n$ so that $\alpha_i = \alpha_0$ and $\lambda_i = \lambda_0$ with $\alpha_0 = h(\gamma_0)$, $\lambda_0 = h(\delta_0)$, $\gamma_0, \delta_0 \in \mathbb{R}$ ($d = r = 1$). In practice, we may also allow empty $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ ($d = r = 0$), i.e., α_i and λ_i have fixed and known values, e.g. $\alpha_i = 0$ and $\lambda_i = 1$ for all $i = 1, 2, \cdots, n$.

As motivated in Section 2.1, the additional covariates making α_i and/or λ_i vary across units should be discrete, i.e. each $\mathbf{U}_i$ and $\mathbf{V}_i$ should be binary/contrast matrices (obtained by an appropriate coding of some categorical predictors) for the MSB model (8). As such, categorical predictors describing site conditions should be included into the design matrices $\mathbf{U}$ and $\mathbf{V}$ to define multiplicative intercepts for groups. Any stage of the considered multistage process can also include categorical variables. In the presence of categorical predictors, the MSB model becomes a multistage analogue of the analysis of covariance model for a binary response: categorical predictors define groups with different intercepts in the study population [69].

2.3. Conjugate Distribution for Measurement Errors

The possibly multivariable integrals (9) in the MSB model specification can make model estimation a difficult task requiring integral approximation methods such as Gaussian quadrature [70,71], or EM-type algorithm [72,73] which can be computationally demanding. To circumvent such difficulties, we consider an implicit *conjugate* distribution for measurement errors, such that each integral (9) is tractable. As per *Proposition 1* below (see a proof in Appendix A.1), under this conjugate distribution of measurement errors, the integral (9) is given by the inverse link function h.

Proposition 1 (Conjugate measurement error distribution)**.** *Assume that the probability distribution corresponding to the link function h in the MSB model (8) has a finite variance $\sigma_o^2 < \infty$. Then, there exists a family of distributions $ME(\boldsymbol{S}_{ij})$ such that the function $\bar{h}$ defined in (9) satisfies*

$$\bar{h}(\boldsymbol{x}_{ij}|\boldsymbol{\beta}_j, \boldsymbol{S}_{ij}) = h\left(\frac{\eta_{ij}}{\sqrt{1 + (\bar{\sigma}_{ij}/\sigma_o)^2}}\right) \tag{11}$$

where $\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j$ and $\bar{\sigma}_{ij}^2 = \boldsymbol{\beta}_j^\top \boldsymbol{S}_{ij} \boldsymbol{\beta}_j$.

Remark 3. *The conjugate ME distribution is known for popular link functions.*

- *For the logit link, $\sigma_o^2 = \frac{\pi^2}{3}$ and the conjugate distribution $ME(\boldsymbol{S}_{ij})$ of measurement errors $\boldsymbol{e}_{ij}$ is the distribution such that the linear combination $\boldsymbol{e}_{ij}^\top \boldsymbol{\beta}_j$ follows a logistic Bridge distribution (which is the conjugate distribution of random-effects in marginalized mixed effects logistic models [74]).*
- *For the probit link, $\sigma_o^2 = 1$ and the conjugate measurement error distribution is the normal distribution with null mean vector and covariance matrix $\boldsymbol{S}_{ij}$. This follows by* Proposition 4 *in [75].*

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

From here on and afterwards, we restrict attention to the measurement error distribution $ME(\boldsymbol{S}_{ij})$ implied by (11), unless otherwise specified. Under the conjugate $ME(\boldsymbol{S}_{ij})$ assumption, the success probability μ_i in (8b) satisfies:

$$\mu_i = \lambda_i \left[\alpha_i + (1-\alpha_i) \prod_{j=1}^{q} h\left(\frac{\eta_{ij}}{\sqrt{1+(\tilde{\sigma}_{ij}/\sigma_o)^2}} \right) \right] \tag{12}$$

where $\lambda_i = h(\boldsymbol{u}_i^\top \boldsymbol{\delta})$, $\alpha_i = h(\boldsymbol{v}_i^\top \boldsymbol{\gamma})$, $\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j$, and $\tilde{\sigma}_{ij}^2 = \boldsymbol{\beta}_j^\top \boldsymbol{S}_{ij} \boldsymbol{\beta}_j$.

2.4. Regularity Conditions and Identifiability

The MSB model (8) is parameterized by the m-vector of all unknown parameters $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top\right)^\top$ with $\boldsymbol{\beta}^\top = \left(\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_q^\top\right)$, $m = p + d + r$. If α_i or λ_i have known values (e.g. $\alpha_i = 0$ and $\lambda_i = 1$), the corresponding sub-vectors ($\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$) are empty ($d = 0$ and $r = 0$, and $\boldsymbol{\theta} = \boldsymbol{\beta}$). Classical inference on $\boldsymbol{\theta}$ is achieved by the ML approach. For inference purposes, we present here the log-likelihood function of the MSB model. Recall that the pmf of a binomial distribution with n_i trials and success probability μ_i has the form [76]

$$P(Y_i = y_i) = \binom{n_i}{y_i} \mu_i^{y_i} (1-\mu_i)^{n_i - y_i}. \tag{13}$$

By the pmf (13), the log-likelihood of a parameter vector $\boldsymbol{\theta}$ given a sample of n independent units $(y_i, \mathbf{X}_i, \mathbf{U}_i, \mathbf{V}_i, \mathbf{S}_i)$ reads

$$\ell_n(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \left[\log \binom{n_i}{y_i} + y_i \log(\mu_i) + (n_i - y_i) \log(1-\mu_i) \right] \tag{14}$$

where μ_i is as defined in Equation (12).

Let us denote $\boldsymbol{\theta}_\circ = \left(\boldsymbol{\beta}_\circ^\top, \boldsymbol{\gamma}_\circ^\top, \boldsymbol{\delta}_\circ^\top\right)^\top$ as the vector of true model parameter. We require the following regularity conditions to ensure the identifiability of the MSB model, i.e., the ability to infer the parameter vector $\boldsymbol{\theta}_\circ$ from observed data.

- **C**1- All predictors in the MSB model are bounded, i.e., there exist compact sets $\mathcal{C}_j \in \mathbb{R}^{p_j}$ (for $j = 1, 2, \cdots, q$), $\mathcal{C}_u \in \mathbb{R}^d$ and $\mathcal{C}_v \in \mathbb{R}^r$ such that $\mathbf{X}_{ij} \in \mathcal{C}_j$, $\mathbf{U}_i \in \mathcal{C}_u$, and $\mathbf{V}_i \in \mathcal{C}_v$ for every $i = 1, 2, \cdots, n$. For every $i = 1, 2, \cdots, n$, $Var[X_{ijk}] > 0$ for $k = 1, 2, \cdots, p_j$, $Var[U_{ik}] > 0$ for $k = 1, 2, \cdots, d$, and $Var[V_{ik}] > 0$ for $k = 1, 2, \cdots, r$. For every $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, q$, and $k = 1, 2, \cdots, p_j$, the X_{ijk} are linearly independent. For every $i = 1, 2, \cdots, n$ and $k = 1, 2, \cdots, d$, and $k = 1, 2, \cdots, d$ the U_{ik} are linearly independent, and for every $i = 1, 2, \cdots, n$ and $k = 1, 2, \cdots, r$, the V_{ik} are linearly independent.
- **C**2- The true parameter vectors $\boldsymbol{\beta}_{j\circ}$, $\boldsymbol{\gamma}_\circ$ and $\boldsymbol{\delta}_\circ$ lie in the interior of known compact sets $\mathcal{B}_j \in \mathbb{R}^{p_j}$ for $j = 1, 2, \cdots, q$, $\mathcal{B}_d \in \mathbb{R}^d$, and $\mathcal{B}_r \in \mathbb{R}^r$ respectively.
- **C**3- There exists, for each stage j, a continuous predictor T_j which is in $\boldsymbol{X}_j$ but not in any $\boldsymbol{X}_k$ (for all $k \neq j$), nor in $\mathbf{U}$, neither in $\mathbf{V}$. Moreover, there exists at least one continuous predictor, either T_u in $\mathbf{U}$ but not in $\mathbf{V}$ nor in $\boldsymbol{X}_j$ (for all $j = 1, 2, \cdots, q$), or T_v in $\mathbf{V}$ but not in $\mathbf{U}$ nor in $\boldsymbol{X}_j$ (for all $j = 1, 2, \cdots, q$).

The conditions **C**1–**C**3 are extensions of the corresponding conditions in Diop et al. [63]. The assumptions **C**1 and **C**2 are classical requirements for identifiability in the TB model. With regard to condition **C**3, it ensures that the product $\prod_{j=1}^{q} \mu_{ij}$ in (8b) is not invariant under any permutation of the parameter vectors $\boldsymbol{\beta}_j$ ($j = 1, \cdots, q$), $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$.

Remark 4. *As emphasized by Diop et al. [63], the condition **C**3 is not particularly stringent.*

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

- *In fact, the motivation of the MSB model assumes that we have some prior knowledge of the steps of the multistage process that generates the binary outcome Y_i. This implies that, for each step that we are aware of, we know about at least one specificity that differentiates it from other known steps. This is reflected by condition **C**3. If our prior knowledge of the steps of the process is confused to the point that two stages share the exact same predictors or risk factors, then we should consider that the two stages are only one known (meaningful) stage.*
- *If all stages of the modeled process are known and accounted for, the model approaches the limiting special case where $\alpha_i = 0$ and $\lambda_i = 1$. In this situation, the distinct continuous predictor in each of the q stages (condition **C**3) is only required for $q - 1$ stages. That is, one of the q stages may not have a distinct continuous predictor.*
- *A continuous predictor is required in the design matrix of either α_i (T_u in $\mathbf{U}$) or λ (T_v in $\mathbf{V}$) only if both α_i* and *λ_i are being estimated. For instance, in the important special cases where either $\alpha_i = 0$ or $\lambda_i = 1$, no distinct continuous predictor is required in any of α_i and λ_i design matrices, if present.*

Based on the conditions **C**1–**C**3, *Proposition* 2 establishes identifiability of the MSB model (8) (see a proof in Appendix A.2).

Proposition 2 (Identifiability of an MSB model)**.** *Under the conditions **C**1–**C**3, the MSB model* (8) *with a logit or probit link function and the conjugate measurement error distribution is identifiable; i.e., $\ell_n(\boldsymbol{\theta}|\boldsymbol{y}) = \ell_n(\boldsymbol{\theta}^*|\boldsymbol{y})$ almost surely implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.*

Remark 5. Proposition 2 *considers the case of conjugate measurement error distribution with known error covariance matrices $\boldsymbol{S}_{ij}$. Note that this includes the special case of no measurement error ($\boldsymbol{S}_{ij} = \mathbf{0}$). Some extensions are also readily available.*

- *For a logit link function,* Proposition 2 *remains true if the conjugate measurement error distribution were replaced by the (multivariate) normal distribution with null mean vector. This follows from Theorem 1 of Shklyar [77], given that the error covariance matrices $\boldsymbol{S}_{ij}$ are assumed known in the MSB model* (8).
- *In the case $q = 1$, the model remains identifiable if all covariance matrices $\boldsymbol{S}_{ij}$ are only known up to a unique scaling factor (see Theorem 2 of Kuchenhoff [78]). In particular, if the unique stage includes only one error-prone continuous predictor, then under conditions **C**1–**C**3, a homogeneous variance of the measurement errors is identifiable along with all other model parameters. Extension to $q > 1$ requires further investigations.*

It is important to observe that a null true model parameter $\boldsymbol{\theta}_0 = \mathbf{0}$ is not generally identifiable (unless $q = 1$, and α_i and λ_i are known constants, which essentially corresponds to the standard binomial model). Indeed, the condition **C**3 implies that some continuous predictors in the design matrices must have non-zero regression slopes. That is, the vector $\boldsymbol{\beta}_j$ of true regression coefficients in each stage j must have at least one non-zero slope.

Also note that along with **C**1 and **C**2, condition **C**3 is sufficient for identifiability of the MSB model (8), but may not be required. That is, identifiability may also be established under some alternative conditions replacing **C**3. These may be physical or biological constraints on the predictors in each stage of the model, or restrictions on model parameters. For instance, in their probit link ovulation-time model, Garel et al. [68] imposed constraints on the parameter space (some coefficients were set to zero) to ensure numerical identifiability.

2.5. Penalized Maximum Likelihood Estimation

The ML estimates of the MSB model (8) parameters can be obtained by solving the score equation $\dot{\ell}_n = \mathbf{0}$. The score function $\dot{\ell}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ell_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}}$ of the MSB model is given in Appendix B. To reduce bias in MSB parameter estimates for small samples and address quasi or complete separation issues, we consider a likelihood penalization approach which is the common route to first order bias correction in the standard logistic regression model [55,60,79–81]. For general exponential family

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

models under canonical parametrization, Firth showed that using Jeffreys invariant prior as likelihood penalty removes the first-order bias term in the asymptotic expansion of the ML estimator [55]. The Jeffreys penalty, which is the square root of the determinant of the Fisher information matrix [82], is asymptotically negligible. As a result, the maximum a posteriori estimator is equivalent to the ML estimator in large samples.

For Penalized Maximum Likelihood (PML) estimation of MSB parameters based on the Jeffreys invariant prior, the penalized log-likelihood function $\tilde{\ell}_n$ is given by

$$\tilde{\ell}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \ell_n(\boldsymbol{\theta}|\boldsymbol{y}) + \frac{1}{2} log|\mathcal{E}_n(\boldsymbol{\theta})| \tag{15}$$

where $log|\cdot|$ returns the logarithm of the determinant of its matrix argument and $\mathcal{E}_n$ denotes the Fisher (expected) information matrix: $\mathcal{E}_n(\boldsymbol{\theta}) = Var\left[\frac{\partial \ell_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right]$ [83]. The PML estimators of the MSB model parameters are obtained by solving the penalized score equation $\dot{\tilde{\ell}}_n = \mathbf{0}$. The Fisher information $\mathcal{E}_n(\boldsymbol{\theta})$ of the MSB model as well as the penalized score $\dot{\tilde{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}}$ are given in Appendix B. For inference in the MSB model, we implemented both the ML and PML procedures in the open source library `msbreg` (available at https://github.com/Chenangnon/msbreg) in R freeware [84].

2.6. Testing for Unit Asymptote

One important feature of the proposed MSB model framework is the ability to infer the absence or not of unknown stages that significantly affect the binomial outcome. Indeed, when there is no controlled experimental condition making the maximum success probability λ_i vary across experimental units, the MSB model includes a multiplicative intercept λ such that $\lambda_i = \lambda$ for all units $i = 1, 2, \cdots, n$ (i.e. the design matrix for λ_i is a vector of all ones: $\mathbf{V} = \boldsymbol{J}_n$). In this situation, the question arises of whether confounding variables significantly affect the binomial outcome at unknown stages.

The absence or presence of unknown stages can be inferred via a statistical test with the null hypothesis H_0: $\lambda = 1$ (unit asymptote). For instance, the nesting relationships between TB and MSB, or ZIB and MSB models allow to always check whether a dataset at hand supports $\lambda = 1$ or not. For $q = 1$, we can test the TB ($\lambda = 1$) model against the more general MSB model ($\lambda < 1$). Likewise, when $q = 2$, we can test the ZIB model (with subject-dependent susceptibility probability and $\lambda = 1$) against the general MSB model ($\lambda < 1$). More generally ($q \geq 1$), testing for a unit asymptote will help achieve a parsimonious fit: when H_0 is rejected, it motivates identifying additional stages (predictors) for the binomial outcome.

Testing the null hypothesis of a unit asymptote H_0 can be achieved, for instance, via an LR test assuming a large sample *chi-square* approximation to the sampling distribution of the LR statistic under H_0. The LR statistic is given by

$$R_\ell\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}_0|\boldsymbol{y}\right) = 2\left[\ell_n(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell_n(\widehat{\boldsymbol{\theta}}_0|\boldsymbol{y})\right] \tag{16}$$

where $\widehat{\boldsymbol{\theta}}$ is the full ML estimate and $\widehat{\boldsymbol{\theta}}_0$ is the ML estimate under the restriction H_0. Since H_0 lies on the boundary of the parameter space ($\lambda = 1$), the naive large sample χ^2_1 approximation to the null distribution of R_ℓ lacks power [85,86]. Indeed, the appropriate large sample null distribution is the so-called *chi-bar-square* distribution $\bar{\chi}^2_{(0,1)} = 0.5 \cdot \chi^2_0 + 0.5 \cdot \chi^2_1$, i.e. the mixture of χ^2_0 (point mass at zero) and χ^2_1 distributions, each with weight 0.5 [86–89]. In practice, this simply implies multiplying the p-value obtained from the χ^2_1 distribution by 0.5.

Under PML inference, the unit asymptote null hypothesis H_0 can be tested using the corresponding penalized LR statistic $R_\ell\left(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_0|\boldsymbol{y}\right)$ where $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}_0$ are respectively the full and restricted PML estimates of $\boldsymbol{\theta}$. Note that because the Jeffreys prior considered for PML inference is asymptotically negligible, the LR statistic remains valid for testing H_0 in large samples under either ML or PML estimation [90]. In small samples ($n < 100$), one can consider a bootstrap approximation of the null distribution of the LR statistic under H_0 (see e.g. [91]).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

3. Simulation Studies

In this section, we showcase the value of the proposed MSB model, starting with simulated case studies, and then exploring important statistical aspects including bias and accuracy, sample size determination, confidence interval and significance of model parameters. For demonstration purposes, we consider a simulated biological case study where the interest is in the viability of a mutant protein resulting from a single amino acid substitution. The outcome variable Y, which is binary (viable/unviable), depends upon the protein's ability to fold into a native conformation, and its ability to bind other subunits or ligands. Free energy changes ($\Delta\Delta G_{fold}$, denoted X_1 and $\Delta\Delta G_{bind}$, denoted X_2) indicate the extent to which a mutation destabilizes protein folding and binding, and we model how they affect the outcome Y.

3.1. Simulated Examples on Mutated Protein Viability

3.1.1. Example 1: Simple Logistic Regression Analysis

For simplicity in the presentation, we start with a model having only one observed predictor: free energy of folding (X_1). We use one simulated dataset (sample size $n = 300$) generated from the logit link model:

$$\begin{aligned} & Y \overset{ind}{\sim} \mathcal{B}er(\mu) \\ \text{where} \quad & \mu = \lambda \cdot h(\beta_0 + \beta_1 X_1), \end{aligned} \tag{17}$$

h is the inverse logit link function defined in (3), $\beta_0 = 3$, $\beta_1 = -2$, X_1 is generated from $\mathcal{U}nif(-1, 4)$ (i.e. a uniform distribution on $[-1, 4]$), and $\lambda = 0.75$. The parameter $\lambda < 1$ means that there are unmeasured variables (e.g. X_2) that affect the protein viability curve.

The scatter plot of the simulated data and the underlying true viability curve are displayed on Figure 1. The traditional approach in the face of the dataset is to fit a Standard Logistic Regression (SLR) model to estimate the parameters β_0 and β_1, to obtain the viability curve as a function of the predictor $\Delta\Delta G_{fold}$. This fit results in the biased estimates $\hat{\beta}_0 = 0.93$ and $\hat{\beta}_1 = -1.14$ (Table 1). The corresponding fitted viability curve (Figure 1) shows large discrepancies with the true viability curve.

The SLR model fit wrongly suggests for instance that the protein viability can approach one as the folding stability $\Delta\Delta G_{fold}$ decreases below $X_1 = -1$ $kcal/mol$ (whereas the true viability plateaus at $\lambda = 0.75$). The fit also incorrectly indicates that varying the folding stability $\Delta\Delta G_{fold}$ from -1 to 0 $kcal/mol$ will result into important changes in the protein viability whereas the true viability is nearly constant in this region. Clearly, mis-specifying the SLR model for the Multistage Logistic Regression (MLR) model can lead to biased estimates of regression parameters and probability curve, and ultimately to incorrect biological interpretations.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

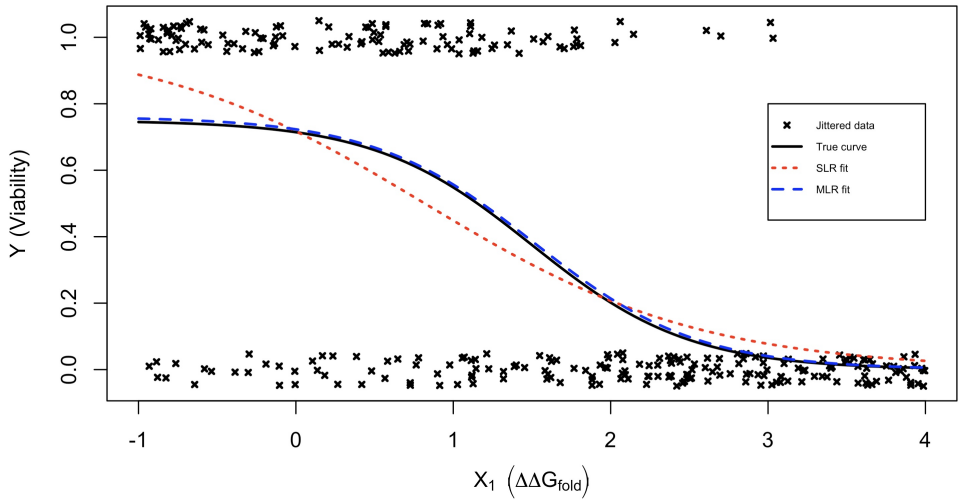


Figure 1. Scatter plot of the simulated working data points (with jitters to improve readability) as a function of the predictor X_1 ($\Delta\Delta G_{fold}$ $[kcal/mol]$), the true viability curve, and the fitted curves using the Standard Logistic Regression (SLR) and the Multistage Logistic Regression (MLR) models.

From both our understanding of protein viability and the scatter in Figure 1, we know or at least suspect that the expected protein viability does not approach one as X_1 decreases continuously below $X_1 = -1\ kcal/mol$. The MLR model approach can thus yield more realistic estimates. We note that the MLR regression parameter estimates shown in Table 1 are closer to the true parameter values, $\hat{\beta}_0 = 2.95$ and $\hat{\beta}_1 = -1.94$. Along with the estimate of the asymptote $\hat{\lambda} = 0.76$, these estimates produce a fitted curve showing only mild discrepancies with the true viability curve (Figure 1). We can measure the overall discrepancy between the fitted and true curves using the mean absolute deviation between the predicted viability and the true viability, $mad(\hat{\mu})$, and their squared Pearson correlation, $R^2(\hat{\mu})$. We find $mad(\hat{\mu}) = 0.01$ and $R^2(\hat{\mu}) = 1.00$ for the MLR model fit, whereas the SLR model fit gives $mad(\hat{\mu}) = 0.06$ and $R^2(\hat{\mu}) = 0.95$. This indicates that the MLR model predicted viability is on average 1% below or above the true viability, versus 6% for the SLR model predictions. By allowing one more parameter (λ), the MLR model provides a better fit to the data.

In practice, the true viability curve is unknown, hence $mad(\hat{\mu})$ or $R^2(\hat{\mu})$ could not be computed. However, the Akaike's Information Criterion (AIC) is smaller for the MLR model fit (AIC = 276.30) as compared to the SLR model fit (AIC = 281.43), indicating that the MLR model fit is better. The estimate of the asymptote ($\hat{\lambda} = 0.76$) has the approximate 95% confidence interval $CI_{95\%} = [0.63, 0.86]$. The LR test comparing the fits of the MLR model and the SLR model (which is nested into the MLR model) indicates with 95% confidence that the improvement achieved by estimating one additional parameter in the MLR model is significant ($\bar{\chi}^2_{(0,1)} = 7.16$, $P = 0.0037$) and is thus worth the lost of one degree of freedom. Along with the AIC statistic and the 95% confidence interval for λ, this test provides evidence for $\lambda < 1$, suggesting there are important but missing stages or covariates. Indeed, one advantage of the MLR model framework over the SLR model is the ability to infer or rule out the absence of important stages or covariates along the process generating the data.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Table 1. Simulated example 1 ($n = 300$): simple logistic regression analysis for protein viability against folding stability ($\Delta\Delta G_{fold}$): comparison of Standard Logistic Regression (SLR) and Multistage Logistic Regression (MLR) model fits.

Parameter	SLR		MLR	
	Estimate (SE)	$CI_{95\%}$	**Estimate (SE)**	$CI_{95\%}$
β_0	0.9300 (0.2031)	[0.5319, 1.3281]	2.9461 (0.8613)	[1.2581, 4.6342]
$\beta_1(\Delta\Delta G_{fold})$	-1.1359 (0.1302)	[-1.3911, -0.8806]	-1.9427 (0.3915)	[-2.7100, -1.1755]
δ	-	-	1.1586 (0.3262)	[0.5192, 1.7979]
$\lambda = h(\delta)$	-	-	0.7611 (0.0593)	[0.6269, 0.8579]
Deviance	277.4280		270.3012	
AIC	281.4280		276.3012	
R^2_{dev}	0.3054		0.3232	

Table notes: SE are standard errors; $CI_{95\%}$ refers to 95% confidence interval determined using the large sample normal approximation to the distribution of each estimate; λ (not a free model parameter) is computed from δ; Deviance refers to the residual deviance of a model fit as compared to a null (constant) fit; R^2_{dev} is the percent deviance reduction achieved by including the predictor X_1 ($\Delta\Delta G_{fold}$) in the fit [92]. Likelihood ratio test for the additional asymptote parameter λ in the MLR fit: $\tilde{\chi}^2_{(0,1)} = 7.1630$, $P = 0.0037$.

Since the results discussed above are based on a unique dataset simulated from a MLR setting, they do not indicate that the MLR fit is always better than the SLR fit, even when the data comes from a MLR process. They nevertheless suggest that one always ponder the SLR model assumption which requires the success probability of the response to arbitrarily approach one under some conditions only dependent upon measured covariates. If the physical or biological process under study may not agree with such an assumption, it is then worthwhile comparing the SLR model and MLR fits and selecting the best model using an objective statistical procedure (e.g. AIC, LR test).

In order to make comparisons robust to randomness in a single dataset, we computed the relative bias, defined as

$$bias(\hat{\beta}_j) = 100 \times \frac{E[\hat{\beta}_j] - \beta_j}{|\beta_j|}$$

and the accuracy, measured by the relative root mean square error,

$$rmse(\hat{\beta}_j) = 100 \times \frac{\sqrt{E[(\hat{\beta}_1 - \beta_1)^2]}}{|\beta_j|}.$$

The expectations in the *bias* and *rmse* formulae are approximated using averages over Monte Carlo replicates of the dateset. To ensure that the standard error of the approximated *bias* and *rmse* are less than 0.5%, we estimated the minimum number of Monte Carlo repetitions to be $95 \times (100 - 95)/(0.5^2) = 1900$ [93]. We thus used 2000 dateset replicates.

As it is expected under a MLR data model, the MLR estimates are less biased and more accurate than logistic estimates, resulting in smaller discrepancies between true and fitted viability curves for the MLR fits. Indeed, replicating the working example 2000 times, we found $bias(\hat{\beta}_0) = 3\%$, $rmse(\hat{\beta}_0) = 30\%$, $bias(\hat{\beta}_1) = -2\%$, and $rmse(\hat{\beta}_1) = 21\%$ for the MLR regression coefficient estimators. The corresponding performance measures for the SLR model estimators are $bias(\hat{\beta}_0) = -71\%$, $rmse(\hat{\beta}_0) = 71\%$, $bias(\hat{\beta}_1) = 4\%$, and $rmse(\hat{\beta}_1) = 44.28\%$, showing a switch of the directions of the biases and larger absolute value of *bias* and *rrmse*. In addition, the expected *mad* and R^2 are $mad(\hat{\mu}) = 0.03$ and $R^2(\hat{\mu}) = 0.99$ for the MLR fits, versus $mad(\hat{\mu}) = 0.06$ and $R^2(\hat{\mu}) = 0.94$ for the SLR fits.

3.1.2. Example 2: Two-Stage Logistic Regression Analysis

The regression analysis in *Example 1* is based on only one predictor. We consider here a model for protein viability with free energy of folding (X_1) and binding (X_2). In this case, the MSB model

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

does not reduce to any other known binomial model. For demonstration, we again use one simulated dataset ($n = 300$) generated from the MLR model:

$$\begin{aligned} & Y \overset{ind}{\sim} \mathcal{B}er(\mu) \\ \text{where} \quad & \mu = \lambda \cdot h(\beta_{01} + \beta_1 X_1) \cdot h(\beta_{02} + \beta_2 X_2), \end{aligned}$$

$\beta_{01} = \beta_{02} = 3$, $\beta_1 = -2$, $\beta_2 = -1$, $X_1 \sim \mathcal{U}nif(-1, 4)$ and $X_2 \sim \mathcal{U}nif(-3, 7)$ are independent, and $\lambda = 0.80$. Here, $\lambda < 1$ stands for further (unmeasured) variables (e.g. residue sites) that affect the viability of the mutated protein.

Fitting a traditional logistic regression model (i.e. two predictors with additive-effects) gives $\hat{\beta}_0 = 0.68$ ($CI_{95\%} = [0.25, 1.12]$) and $\hat{\beta}_1 = -0.97$ ($CI_{95\%} = [-1.23, -0.70]$) and $\hat{\beta}_2 = -0.36$ ($CI_{95\%} = [-0.48, -0.25]$) with AIC $= 259.80$. This fit results in predicted success probabilities $\hat{\mu}$ with mean absolute deviation from the truth $mad(\hat{\mu}) = 0.07$ and predictive power $R^2(\hat{\mu}) = 0.89$. Under a multiplicative-effect assumption based on the biology of protein viability, the fitted Zero-Inflated Logistic (ZIL) and MLR models are shown in Table 2. As expected, the ZIL and MLR models provide better fits to the data (higher deviance reduction ratio R^2_{dev}, lower AIC) as compared to the traditional logistic fit. It appears that the MLR model is fitter than the ZIL model based on its lower AIC and higher R^2. The 95% confidence interval for λ ($CI_{95\%} = [0.63, 0.91]$) as well as the LR test for the ZIL fit ($\lambda = 1$) versus the MLR fit ($0 < \lambda < 1$) also point to the choice of the MLR model ($\bar{\chi}^2_{(0,1)} = 14.06$, $P < 0.001$).

Table 2. Simulated example 2 ($n = 300$): bivariate logistic regression analysis for protein viability against folding stability ($\Delta\Delta G_{fold}$) and binding stability ($\Delta\Delta G_{bind}$): comparison of Zero-Inflated Logistic (ZIL) and Multistage Logistic Regression (MLR) model fits.

Parameter	ZIL		MLR	
	Estimate (SE)	$CI_{95\%}$	**Estimate (SE)**	$CI_{95\%}$
β_{01}	1.2107 (0.3296)	[0.5648, 1.8566]	2.9969 (1.1077)	[0.8259, 5.1679]
$\beta_1(\Delta\Delta G_{fold})$	-1.1456 (0.1758)	[-1.4901, -0.8010]	-1.9110 (0.5003)	[-2.8916, -0.9304]
β_{02}	4.2708 (1.4974)	[1.3360, 7.2055]	4.5245 (1.5648)	[1.4575, 7.5914]
$\beta_2(\Delta\Delta G_{bind})$	-1.1949 (0.3353)	[-1.8520, -0.5378]	-1.2602 (0.3542)	[-1.9545, -0.5659]
δ	-	-	1.5059 (0.4904)	[0.5447, 2.4671]
$\lambda = h(\delta)$	-	-	0.8185 (0.0729)	[0.6329, 0.9218]
Deviance	229.5335		222.5019	
AIC	237.5335		232.5019	
R^2_{dev}	0.3678		0.3872	

Table notes: SE are standard errors; $CI_{95\%}$ refers to 95% confidence interval determined using the large sample normal approximation to the distribution of each estimate; λ (not a free model parameter) is computed from δ; Deviance refers to the residual deviance of a model fit as compared to a null (constant) fit; R^2_{dev} is the percent deviance reduction achieved by including the predictors X_1 ($\Delta\Delta G_{fold}$) and X_2 ($\Delta\Delta G_{bind}$) in the fit [92]. Likelihood ratio test for the additional asymptote parameter λ in the MLR fit: $\bar{\chi}^2_{(0,1)} = 14.0632$, $P < 0.001$.

Beyond this one dataset simulation, the MLR model estimates are less biased and more accurate as compared to the ZIL estimates, as expected for data generated from a MLR process. Indeed, from Table 3–which shows the mean estimates, relative biases, root mean square errors of parameter estimates, and predictive power (computed from ZIL and MLR model fits on 2000 replicates of the MLR data)– we notice that the MLR model estimates have much less bias than ZIL estimates, and slightly lower *rmse*.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Table 3. Example 2: comparison of performance measures (mean estimates, relative biases and root mean square errors) for ZIL and MLR model fits using 2000 replicates of MLR data ($n = 300$).

Measure	Model	Model parameter (true value)					Predictive power	
		β_{01} (3)	β_1 (-2)	β_{02} (3)	β_2 (-1)	λ (0.8)	$mad(\hat{\mu})$	$R^2(\hat{\mu})$
Mean	ZIL	2.1830	-1.6863	2.2972	-0.8708	-	0.0333	0.9690
	MLR	3.3368	-2.1564	3.2767	-1.0685	0.8022	0.0280	0.9806
bias (%)	ZIL	-27.23	15.69	-23.43	12.92	-	-	-
	MLR	11.23	-7.82	9.22	-6.85	0.28	-	-
rmse (%)	ZIL	50.42	34.49	45.00	30.40	-	-	-
	MLR	43.18	32.54	40.01	29.35	9.55	-	-

Table notes: The ZIL and MLR models differ by the presence of the additional parameter $\lambda = h(\delta)$ in the MLR model. The ZIL model assumes that $\lambda = 1$. *mad* (mean absolute deviation) is the sample average of the absolute deviations between the predicted success probabilities ($\hat{\mu}$) from a fit and the true success probabilities (μ). R^2 is the squared Pearson correlation coefficient between $\hat{\mu}$ and μ.

3.2. General Performance in Synthetic Data

In the examples discussed in section 3.1, we have considered $\lambda \in \{0.75, 0.80\}$, intercept equal to 3 and slopes in $\{-1, -2\}$, set the sample size to $n = 300$, and used a uniform distribution for the predictors. Our goal here is to assess the behavior of the PML estimators of the MSB model when these parameters vary. We considered regression intercepts in $\{3, 8\}$ and slopes in $\{-1, -2, -3\}$, and the predictor distributions (*PD*) include normal, and log-normal in addition to uniform. We generally investigated sample sizes $n \in [100, 2000]$ and asymptote values $\lambda \in [0.5, 1]$, although we occasionally used n up to $n = 10000$ and λ down to $\lambda = 0.25$ to illustrate some specific trends.

We summarize the key results on the performance of PML inference in simulated MSB data. We consider aspects such as the bias and accuracy of the PML estimators, their implication for sample size determination under an MSB design, the ability of the LR test procedure to detect the absence of unknown stages ($\lambda < 1$), and the coverage and significance probabilities of 95% confidence interval for a regression slope. Detailed simulation design and results are available in Appendix C.

3.2.1. Bias and Accuracy

Figures 2 and 3 show, respectively, the relative *bias* and the *rmse* of the PML estimates of λ, β_0 and β_1 as functions of the true asymptote λ, the sample size n, and the *PD* using data from the one stage model (17) with $\beta_0 = 3$ and $\beta_1 = -2$. It appears that for the estimates $\hat{\lambda}$, $\hat{\beta}_0$ and $\hat{\beta}_1$, both the relative *bias* and *rmse* decrease and approach zero as n increases

Bias in $\hat{\lambda}$. At $\lambda = 1$, the relative *bias* of $\hat{\lambda}$ is close to zero for all considered *PD* and sample sizes $n \in [100, 2000]$. As λ decreases, the estimate $\hat{\lambda}$ exhibits a slight *bias*, reaching a maximum of about 10% in absolute value at $n = 100$ when $\lambda = 0.5$ (Figure 2). We notice a pattern in the *bias*: under a uniform *PD* for instance, the *bias* is negative at $\lambda = 0.9$ and positive for $\lambda \in \{0.5, 0.75\}$. Additional simulations with λ values down to 0.25 (see Appendix C) reveal that under a uniform or log-normal *PD*, the *bias* in $\hat{\lambda}$ is negative (under-estimation) for high λ values ($0.75 < \lambda < 1$), essentially zero for medium λ values ($0.5 \leq \lambda \leq 0.75$), and positive (over-estimation) for low λ values ($0.25 \leq \lambda < 0.5$). For the normal *PD*, the pattern persists, but the *bias* is negative for $0.8 < \lambda < 1$.

Accuracy of $\hat{\lambda}$. At $\lambda = 1$, the relative *rmse* of $\hat{\lambda}$ is close to zero for all *PD* and sample sizes. The accuracy also decreases with λ (Figure 3). Indeed, under the uniform *PD* with $n = 100$, the *rmse* of $\hat{\lambda}$ is about 6% for $\lambda = 0.9$, but reaches about 21% for $\lambda = 0.5$. We observe a similar trend under the log-normal *PD*, with a *rmse* around 5% at $\lambda = 0.9$, and 19% at $\lambda = 0.5$ ($n = 100$). The estimate $\hat{\lambda}$ is comparatively less accurate under a normal *PD*, with a *rmse* around 10% at $\lambda = 0.9$, and 25% at $\lambda = 0.5$ ($n = 100$).

Bias and accuracy of $\hat{\beta}_0$ and $\hat{\beta}_1$. At $\lambda = 1$ (all *PD*), the *bias* of regression parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ is also close to zero (Figure 2), and the *rmse* (which is then essentially the standard deviation for each estimator) is about 15% at $n = 300$, 10% at $n = 500$, and 5% at $n = 1500$ (Figure 3). As λ decreases, the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

relative *rmse* increases, reaching 32% for $\hat{\beta}_0$ and 22% for $\hat{\beta}_1$ when $\lambda = 0.5$ and $n = 500$. This indicates that the lower λ, the larger the sample size required to achieve the same level of accuracy for regression coefficients.

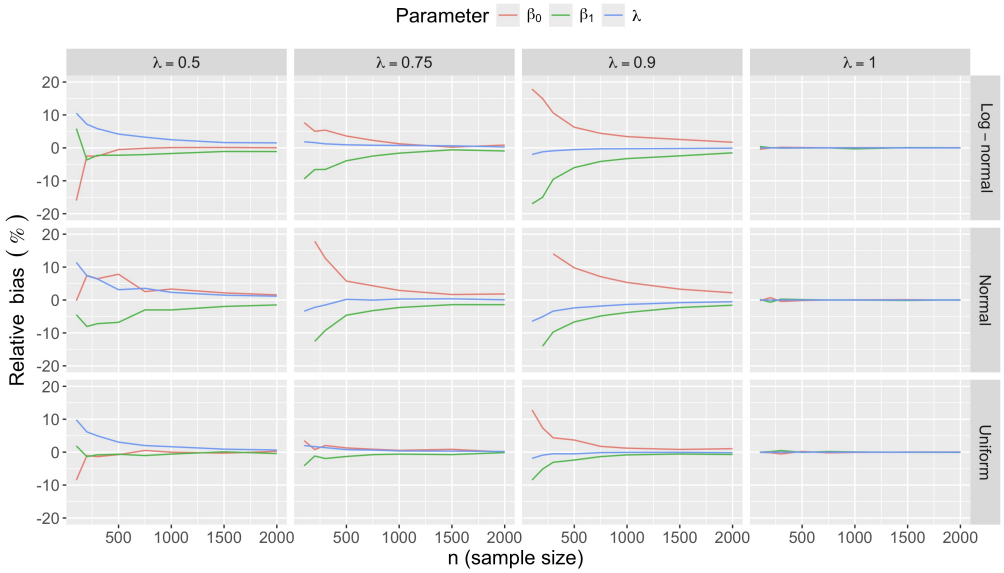


Figure 2. Relative bias (*bias*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB data generation process in Equation (17) with $\beta_0 = 3$, $\beta_1 = -2$. The figure columns show the investigated true λ values, and rows correspond to investigated distributions of predictor.

Effect of predictor distribution. The distribution of the predictor does not affect the accuracy of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for λ close to one (Figure 3). But as λ decreases, the relative *rmse* is *ceteris paribus* minimum under uniform *PD* and maximum under log-normal *PD* for both $\hat{\beta}_0$ and $\hat{\beta}_1$. At $\lambda = 0.5$ for instance, $rmse(\hat{\beta}_1) = 15\%$ at $n = 1000$ under a uniform predictor. In contrast, $rmse(\hat{\beta}_1) = 18\%$ at $n = 2000$ under a log-normal predictor.

Impact of regression coefficients. The trends in the relative *bias* and *rmse* shown for $\beta_0 = 3$ and $\beta_1 = -2$ (Figures 2 and 3) are similar across combinations of $\beta_0 \in \{3, 8\}$ and $\beta_1 \in \{-1, -2, -3\}$. Specifically, the *rrmse* of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ increase with λ, and for $\lambda < 1$, the *rrmse* is *ceteris paribus* minimum under uniform *PD* and maximum under log-normal *PD* (see Appendix C).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

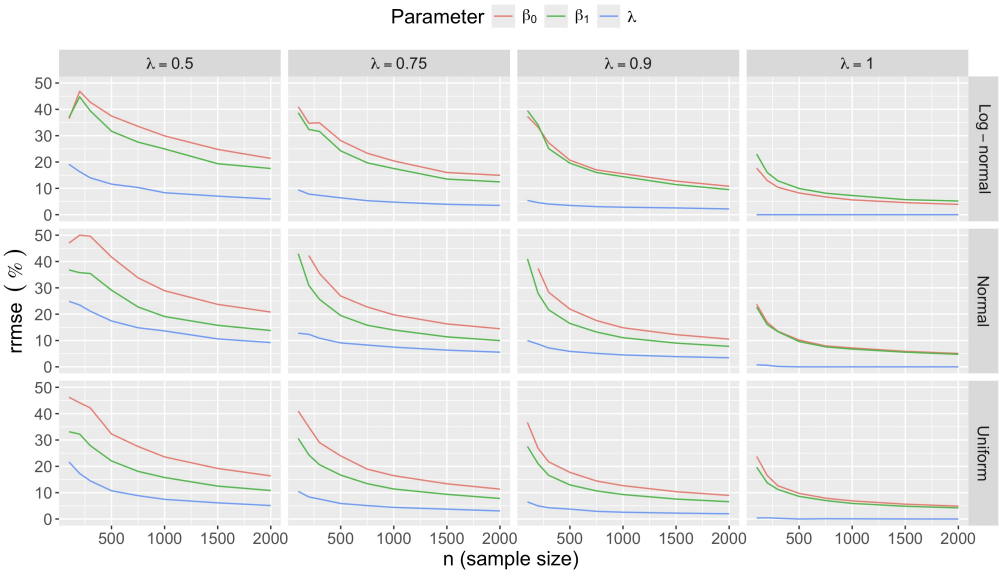


Figure 3. Relative root mean square error (*rrmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB data generation process in Equation (17) with $\beta_0 = 3$, $\beta_1 = -2$. The figure columns show the investigated true λ values, and rows correspond to investigated distributions of predictor.

Beta regression on the accuracy of $\hat{\beta}_1$. To summarize the general trends for the slope estimate $\hat{\beta}_1$, we fitted a beta regression to the values of the relative *rmse* of $\hat{\beta}_1$ as a function of the simulation factors. Table 4 presents the results of the fit which explain about 87% of the variability in $rmse(\hat{\beta}_1)$. From Table 4, we observe that the leading factors in the variations of $rmse(\hat{\beta}_1)$ are λ, *PD* and β_0, with significant interactions. For instance, decreasing λ by 0.1 results on average in a 6.5% increase in $rmse(\hat{\beta}_1)$ under a uniform *PD* and a small sample ($n < 200$). Switching then to a log-normal *PD*, we have on average a 16% increase in $rmse(\hat{\beta}_1)$ for small λ values, and about a 26% increase for λ values close to one. Under scenarios with $\beta_0 = 8$, we observe an average 7% decrease in $rmse(\hat{\beta}_1)$ as compared with $\beta_0 = 3$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Table 4. Beta regression model fit to the relative root mean square error ($rmse/100$) of the estimate $\hat{\beta}_1$ of the slope in a simple logistic regression with a multiplicative intercept λ.

Term	Mean $rmse/100$ (log link)		Precision parameter ϕ (log link)	
	Estimate (SE)	$CI_{95\%}$	**Estimate (SE)**	$CI_{95\%}$
Intercept	-0.8591 (0.0174)	[-0.8931, -0.8251]	6.1178 (0.2054)	[5.7152, 6.5204]
n	-0.0003 (0.0000)	[-0.0003, -0.0003]	0.0020 (0.0002)	[0.0016, 0.0024]
λ	-0.6482 (0.0278)	[-0.7027, -0.5937]	-2.3751 (0.2475)	[-2.8602, -1.8899]
$\beta_0 = 8$	-0.0688 (0.0104)	[-0.0892, -0.0484]	0.2182 (0.0741)	[0.0730, 0.3633]
β_1	-0.0026 (0.0027)	[-0.0078, 0.0026]	-	-
PD = Normal	0.0942 (0.0190)	[0.0570, 0.1314]	-0.7010 (0.1460)	[-0.9872, -0.4149]
PD = Log-normal	0.1588 (0.0190)	[0.1216, 0.1960]	-0.6797 (0.1460)	[-0.9658, -0.3935]
$n \times \lambda$	-0.0006 (0.0000)	[-0.0006, -0.0005]	-0.0012 (0.0002)	[-0.0016, -0.0007]
$n \times (\beta_0 = 8)$	0.0001 (0.0000)	[0.0001, 0.0001]	-	-
$n \times (PD = \text{Normal})$	0.0001 (0.0001)	[0.0001, 0.0001]	0.0006 (0.0001)	[0.0003, 0.0009]
$n \times (PD = \text{Log-normal})$	0.0001 (0.0001)	[0.0001, 0.0002]	0.0003 (0.0001)	[0.0001, 0.0006]
$\lambda \times (PD = \text{Normal})$	0.0499 (0.0293)	[-0.0075, 0.1072]	-	-
$\lambda \times (PD = \text{Log-normal})$	0.1010 (0.0295)	[0.0432, 0.1588]	-	-

Table notes: the predictors include the sample size ($n \in [100, 2000]$), the asymptote parameter ($\lambda \in [0.25, 1]$), the intercept ($\beta_0 \in \{3, 8\}$) and the slope ($\beta_1 \in \{-3, -2, -1\}$) in the data model for simulations. The dependent variable in this beta regression fit is $rmse/100$ where $rmse$ is the relative root mean square error of $\hat{\beta}_1$ expressed in percent of the true β_1 value (so $rmse/100 \in (0, 1)$). The $rmse$ values were computed from 2000 Monte Carlo replicates of each simulation setting. The precision parameter ϕ is such that, if the mean of $rmse/100$ is denoted μ, the variance of $rmse/100$ is $\sigma^2 = \mu(1-\mu)/(1+\phi)$. The beta regression model was fitted using the R package *betareg* [94]. The model was initially fitted using all predictors and all two-order interactions as terms in the linear predictors of both μ and σ^2. Then, a backward selection was used to recursively drop the least important term based on Akaike's Information Criterion (AIC), but main terms (predictors) were not dropped from the linear predictors of the mean $rmse/100$. SE denotes standard errors; $CI_{95\%}$ refers to 95% confidence interval determined using the large sample normal approximation to the distribution of each estimate; $AIC = -6342.84$, $R^2 = 0.87$ (squared correlation between linear predictor and link-transformed response).

The presence of a significant negative interaction between n and λ (Table 4) indicates that larger sample sizes are generally required when $\lambda < 1$ as compared to when $\lambda = 1$. Indeed, the $rmse$ generally decreases with n, but the interactive effect of λ with the sample size n is also negative and relatively large (twice the main effect of n) as compared to the marginal effect of n (Table 4). For instance, increasing n by 100 would result in a 5.5% average decrease in $rmse(\hat{\beta}_1)$ if $\lambda = 0.5$, versus a 8.3% average decrease if $\lambda = 1$.

3.2.2. Sample Size Determination

The trends in the accuracy of PML estimates indicate that the sample size in an MSB design depends on λ. To illustrate the importance of such trends for sample size calculation when considering an MSB framework for data analysis, we run additional simulations with sample sizes up to $n = 10000$ under a uniform PD with true slope $\beta_1 = -2$ and intercept $\beta_0 = 3$. Figure 4 shows for different λ values, the sample size n required to achieve a specified root mean square error for the estimate $\hat{\beta}_1$.

It appears that for $\lambda = 1$, a sample size of about $n = 393$ is required to achieve $rmse(\hat{\beta}_1) = 10\%$. Reaching the same level of accuracy with $\lambda = 0.9$ requires a sample of size $n = 853$. The required sample size becomes $n = 1313$ for $\lambda = 0.75$, and reaches $n = 2417$ when $\lambda = 0.5$. Hence, in this particular simulation setting, halving λ demands about six times more samples to achieve the same 10% level of accuracy. To achieve $rmse(\hat{\beta}_1) = 5\%$, $n = 1425$ samples are required when $\lambda = 1$, $n = 3609$ when $\lambda = 0.9$, $n = 5508$ when $\lambda = 0.75$ and $n = 9949$ when $\lambda = 0.5$. Hence for a 5% level of accuracy, halving λ demands about seven times more samples.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

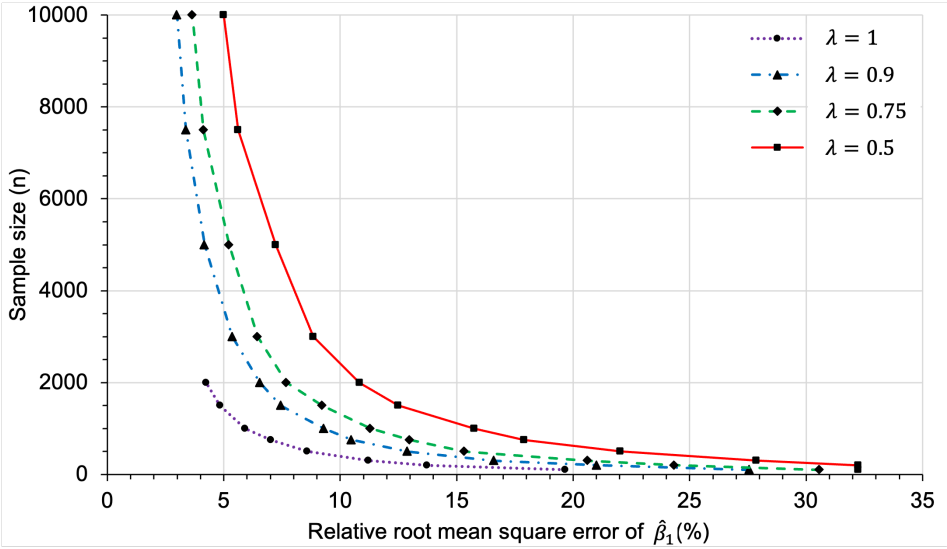


Figure 4. Sample size n required to achieve a specified relative root mean square error for the Penalized Maximum Likelihood estimate $\hat{\beta}_1$ of the regression slope β_1 under the MSB data generation process in Equation (17) with a uniform predictor distribution, $\beta_0 = 3$, $\beta_1 = -2$ and selected λ values.

We observe similar trends under normal and log-normal *PD* (see details in Appendix C). For instance, under the normal *PD*, to achieve $rrmse(\hat{\beta}_1) = 10\%$, about $n = 480$ samples are required when $\lambda = 1$. With $\lambda = 0.5$, the required sample size becomes $n = 3819$, that is, almost eight fold increase of the sample size requirement. Under the log-normal *PD*, about $n = 497$ samples are required to achieve $rrmse(\hat{\beta}_1) = 10\%$ when $\lambda = 1$. With $\lambda = 0.5$, the required sample size becomes $n = 5779$, that is, 11.6 fold increase of the sample size requirement.

3.2.3. Testing for Unit Asymptote

Figure 5 shows the rejection rate of the LR test of the null hypothesis ($H_0 : \lambda = 1$) at the significance level of 5% for true λ values in $[0.5, 1]$ and regression slopes $\beta_1 = -3, -1$. It appears that when the TB model is the truth ($\lambda = 1$) and the sample size is large ($n = 2000$), the rejection rate of the LR test is close to the nominal level of 5% (5.5% for uniform *PD*, 4.5% for normal *PD* and 4.9% for log-normal *PD*). This shows that the LR test detects the TB model when it is the correct model and sample size is large. Interestingly, this remains true in small sample with rejection rates under 6.7% (largest value observed under log-normal *PD* with $\beta_1 = -3$).

For true $\lambda \leq 0.9$, we observed on Figure 5 that the power of the LR test is 1 under uniform *PD* in samples of size $n \geq 1000$. In smaller samples, the power decreases with decreasing sample size, but increases as the true λ decreases, reaching 59% at $\lambda = 0.5$. In other words, the LR test becomes conservative in small samples. We also notice that the rejection rates are lower under normal and log-normal *PD* as compared to the uniform *PD*. Finally, the rejection rates are lower for $\beta_1 = -3$ as compared to $\beta_1 = -1$ when the *PD* deviates from uniform.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

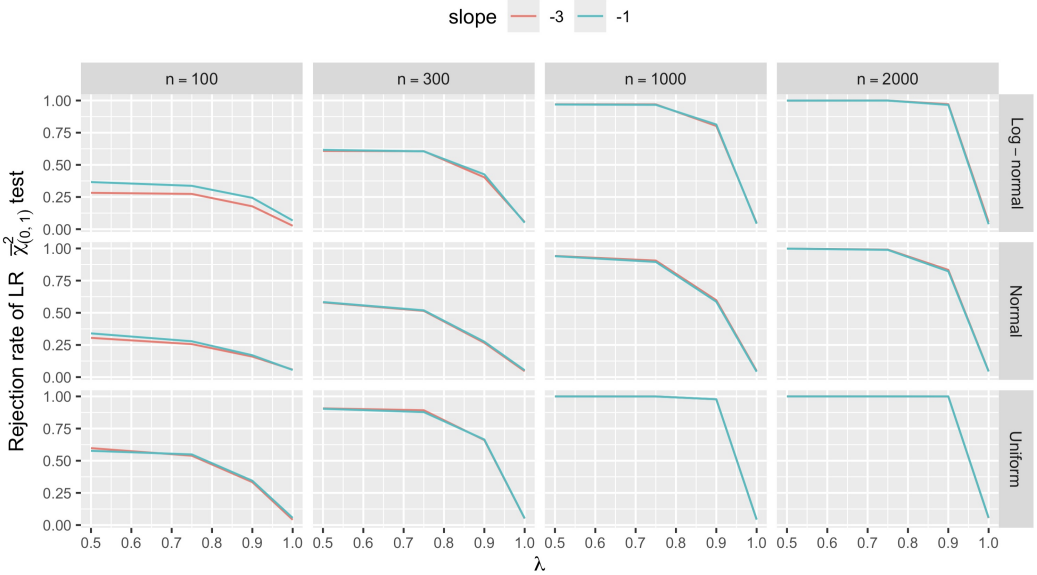


Figure 5. Rejection rate of the Likelihood Ratio (LR) test of the null hypothesis (H_0: $\lambda = 1$) at the significance level of 5% under the MSB model data generation process in Equation (17) with $\lambda \in [0.5, 1]$, $\beta_0 = 3$, and $\beta_1 \in \{-1, -3\}$.

3.2.4. Statistical Inference on Regression Coefficients

We present the finite sample properties of inference based on the large sample normal approximation of the distributions of PML estimates under the MSB model. Figure 6 shows the Coverage Probability (CP) of approximate 95% confidence interval using the inverse negative Fisher information matrix as covariance matrix for PML estimates. We observed that the CP is close to the nominal level (95%) and stays above 90% when either $\lambda = 1$, or the sample size is large, or the predictor distribution is uniform or normal. For a log-normal predictor however, the CP decreases with λ, reaching 88% at $n = 1000$ for $\lambda = 0.5$.

Figure 7 displays the Significance Probability (SP) of the approximate 95% confidence interval. It appears that, irrespective of the *PD* and λ values, the significance of the slope β_1 is close to 1 for any $n \geq 200$. The intercept coefficient β_0 is more sensitive to both λ and sample size, the SP decreasing and reaching 0.22 under normal *PD* at $n = 100$.

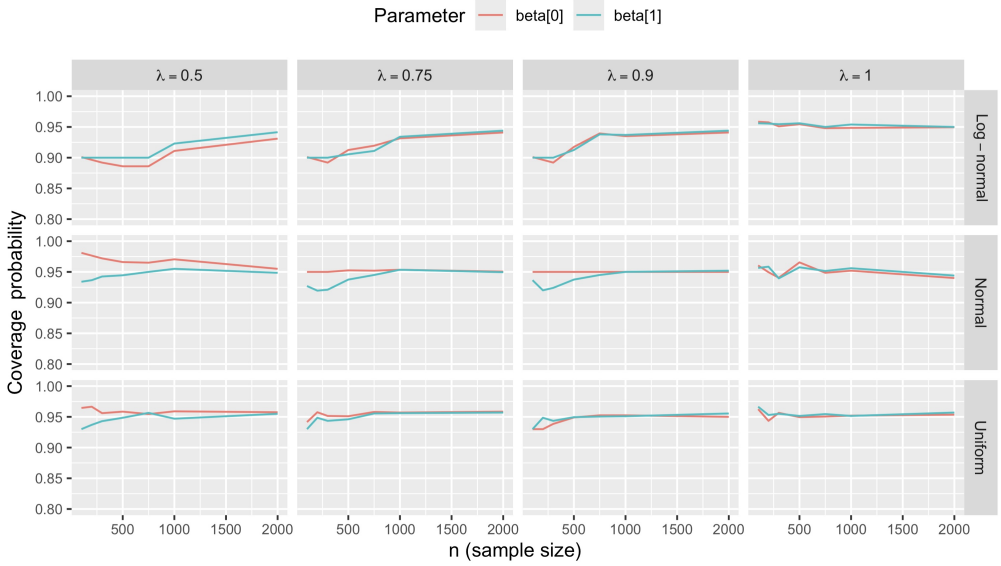


Figure 6. Coverage probability of Penalized Maximum Likelihood estimates of multistage binomial model parameters β_0 (shown as beta[0]) and β_1 (shown as beta[1]) under the MSB data generation process in Equation (17) with $\beta_0 = 3$, $\beta_1 = -1$. The figure columns show the investigated true λ values, and rows correspond to investigated distributions of predictor.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

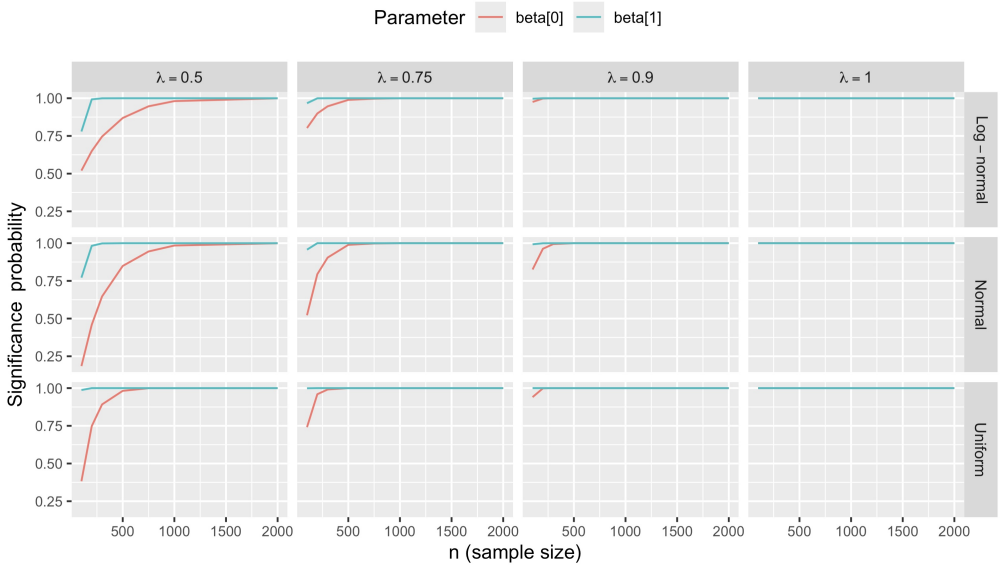


Figure 7. Significance probability of Penalized Maximum Likelihood estimates of multistage binomial model parameters β_0 (shown as beta[0]) and β_1 (shown as beta[1]) under the MSB data generation process in Equation (17) with $\beta_0 = 3$, $\beta_1 = -1$. The figure columns show the investigated true λ values, and rows correspond to investigated distributions of predictor.

4. Discussion

In this paper, we propose a binomial regression framework for modeling binary data arising from processes with multiplicative steps. The use of the model is prescribed when modeling a complex binary process with no data on the detailed discrete stages of the phenomenon. As soon as we have data on each specific stage, a multivariate binomial GLM [95,96] becomes more appropriate for the data analysis. We derived conditions for the identifiability of model parameters. Not surprisingly, it is sufficient that each stage has a continuous predictor only present at that stage to ensure identifiability. The maximum likelihood estimates of the MSB model parameters are asymptotically unbiased, but subject to bias in small samples. To reduce the small sample while addressing separation issues, we developed a penalized maximum likelihood estimation procedure based on Jeffreys invariant prior. We have demonstrated the use context and usefulness of the modeling framework using synthetic data on mutant protein viability (see Figure 1).

Interestingly, both the standard binomial GLM and the ZIB model are special instances of the MSB model. Indeed, though it is not usually described as based on a multiplicative risk process, the ZIB model is based on (2), but limited to $q = 2$. It is apparent from (2) that the overall success probability μ of the binary outcome can be very small. It follows that data generated from an MSB model may have a larger occurrence of zeros as compared to the corresponding standard binomial model. This observation immediately points to the ZIB model which has become popular for modeling binary data with an excess of zeros relative to the expectation under the standard binomial model [63–66]. The ZIB model stems from a mixture of two distinct subpopulations where one subpopulation is subject to a risk factor (group of "susceptible" units) whereas the other subpopulation is not (group of "immune" units). This can be mathematically described as a sequence of two Bernoulli processes. A first Bernoulli process assigns each individual to one of the two subpopulations. Then, the susceptible subpopulation, when exposed, is subject to a risk through a second Bernoulli process, while the immune subpopulation is not subject to any risk, even if exposed. The first stage of the process is considered as a source of nuisance and is not of interest to the data analysis [63]. It is here important to notice that the MSB model (8) is not a zero-inflation model. The model simply restricts the success probability of the binomial outcome Y_i to the subset $(\alpha_i\lambda_i, \lambda_i)$ of $(0, 1)$, equally allowing for inflation of both zeros and ones.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

The asymptote parameter λ included in the MSB model (maximum success probability for the binary outcome) can capture unknown and missing stages. In particular, when only one stage is known for the process under investigation, λ provides an estimate of the importance of unknown or unaccounted stages or components in the determination of the binomial outcome. In health data context for instance, $1-\lambda$ can represent a cure fraction [63], or a maximum survival probability. The standard binomial model would over estimate survival or risk if we omit λ, leading to poor prediction (Figure 1), and possibly to poor policy design. We investigated the finite sample properties of a non-standard likelihood ratio test for detecting $\lambda = 1$. The test have shown rejection rates (of the assumption $\lambda = 1$) close to the nominal level when $\lambda = 1$ (Figure 5). The test is generally conservative and will point to $\lambda < 1$ only when the latter is strongly supported by the data. Hence, when studying a phenomenon without a strong justification that the process is additive, it is worthwhile testing the MSB model against the standard binomial model to identify which model is supported by the available data. Rejecting $\lambda = 1$ would suggest the absence of some important stages or covariates.

We investigated the finite-sample properties of the penalized maximum likelihood estimates via simulations. It turns out that λ affects the estimation of regression coefficients. In particular, lower λ values are associated with higher mean square errors in regression coefficient estimates. This implies that larger samples are required for lower λ values (Figure 4). Intuitively, $\lambda < 1$ means more failures in the binary response as compared to $\lambda = 1$. However, unlike the failures under $\lambda = 1$ which provide as much information on regression parameters as successes, the additional failures due to $\lambda < 1$ do not provide any information on regression parameters. Stated otherwise, a lower λ means that each data point is less informative as compared to when $\lambda = 1$. Hence the effective size of the sample contributing to the estimation of regression parameters is only a small proportion of n when λ is very low. We also established that larger sample sizes are required under log-normal distribution as compared to uniform or normal distribution of predictors.

We finally point to a few open questions that deserve attention. First, a rigorous theory of the existence, consistency and asymptotic normality of the (penalized) maximum likelihood estimators of MSB model parameters should be established. A related aspect of interest is the identifiability and inference when predictor dimension is much larger than the sample size. Second, the bias in parameters estimates relates to the particular parameterization of the MSB model. Nevertheless, first order bias corrected estimates can perform substantially better (in terms of accuracy, i.e. mean square error) than crude maximum likelihood estimates [97]. Although the use of the Jeffreys invariant prior as likelihood penalty ensures finite estimates and reduced bias, it does not completely remove first order bias from the maximum likelihood estimates of the MSB model parameters. Further bias reduction is thus possible using recent general purpose methods developed for nonlinear models (see e.g [98]). Third, diagnostic statistics such as influence measures [99] should be developed for real data analysis under the MSB model framework. Fourth, the power of the test for unit asymptote may be improved by implementing the Bartlett correction [100] for the likelihood ratio statistic. Moreover, the score test, an alternative to likelihood ratio test, may offer a faster procedure for model selection since it only requires fitting the null model (i.e. under H_0: $\lambda = 1$) [101]. Last, but not least, we assumed that measurement error standard deviations are known in this work. Future research should investigate to what extent the proposed conjugate measurement error distribution is robust to distribution mis-specification.

Author Contributions: Conceptualization and methodology, C.T., Y.S. and C.M.; software, C.T.; resources, C.M.; writing—original draft preparation, C.T.; writing—review and editing, C.T., Y.S., and C.M.; visualization, C.T.; supervision and funding acquisition C.M.; All authors have read and agreed to the published version of the manuscript.

Funding: Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The authors confirm that the data supporting the findings of this work are available within the article.

Acknowledgments: The authors acknowledge the support of The Institute for Interdisciplinary Data Sciences for high performance computing. They also thank Holly Wichman and other members of the Institute for Modeling Collaboration and Innovation for their ongoing support of interdisciplinary collaborations.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Proof of Propositions

Appendix A.1. Proof of Proposition 1

Proof. Each factor ξ_{ij} in (4b) can be considered as the success probability of the independent Bernoulli variable Y_{ij} associated with stage j given a covariate vector $\boldsymbol{w}_{ij}$, i.e.,

$$(Y_{ij}|\boldsymbol{W}_{ij} = \boldsymbol{w}_{ij}) \overset{ind}{\sim} \mathcal{B}er(\xi_{ij}). \tag{A1}$$

From (A1), the Bernoulli variable Y_{ij} can further be represented by truncating a continuous distribution, e.g. truncated normal, logistic or scale mixture of skew normal variables [102]. In this framework, Y_{ij} has the stochastic representation

$$\begin{aligned}(Y_{ij}|Z_{ij} = z_{ij}, W_{ij} = w_{ij}) &\overset{d}{=} I_{(0,\infty)}(\zeta_{ij} + z_{ij}) \\ (Z_{ij}|W_{ij} = w_{ij}) &\sim H(0, \sigma_o^2)\end{aligned} \tag{A2}$$

where $\overset{d}{=}$ means "equal in distribution to", and H denotes the distribution with cdf h. For the logit link for instance, $H(\eta, \sigma_o^2)$ is the logistic distribution with mean η and variance $\sigma_o^2 = \pi^2/3$; and in the case of the probit link, $H(\eta, \sigma_o^2)$ is the normal distribution with mean η and variance $\sigma_o^2 = 1$. From Equation (6), we have $\boldsymbol{x}_{ij} = \boldsymbol{w}_{ij} + \boldsymbol{e}_{ij}$. It follows that $\zeta_{ij} + z_{ij}$ is a function of $\boldsymbol{e}_{ij}$ as $\zeta_{ij} + z_{ij} = \eta_{ij} + \left(z_{ij} - \boldsymbol{e}_{ij}^\top \boldsymbol{\beta}_j\right)$. It then appears that the integral in (9) can be simplified by choosing the distribution $\mathcal{ME}(\boldsymbol{S}_{ij})$ of the measurement error $\boldsymbol{e}_{ij}$ such that the distribution of $z_{ij} - \boldsymbol{e}_{ij}^\top \boldsymbol{\beta}_j$ is in the same family as $H(0, \sigma_o^2)$. Indeed, since Z_{ij} and $\boldsymbol{e}_{ij}$ are independent, $Z_{ij} - \boldsymbol{e}_{ij}^\top \boldsymbol{\beta}_j$ has variance $s_{ij}^2 = \sigma_o^2 + \bar{\sigma}_{ij}^2$ with $\bar{\sigma}_{ij}^2 = \boldsymbol{\beta}_j^\top \boldsymbol{S}_{ij} \boldsymbol{\beta}_j$, and since both Z_{ij} and $\boldsymbol{e}_{ij}$ have null mean, the representation (A2) becomes

$$\begin{aligned}(Y_{ij}|\overline{T}_{ij} = \bar{t}_{ij}) &\overset{d}{=} I_{(0,\infty)}(\eta_{ij} + \bar{t}_{ij}) \\ \overline{T}_{ij} &\sim H\left(0, \bar{\sigma}_{ij}^2\right).\end{aligned}$$

In light of (A2), the distribution H in the stochastic representation must have variance σ_o^2 in order to write e.g. $\xi_{ij} = h(\zeta_{ij})$. From $\eta_{ij} + \bar{t}_{ij} > 0 \Leftrightarrow \frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}} + \frac{\bar{t}_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}} > 0$, we get by setting $T_{ij} = \frac{\overline{T}_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}$:

$$\begin{aligned}(Y_{ij}|T_{ij} = t_{ij}) &\overset{d}{=} I_{(0,\infty)}\left(\frac{\eta_{ij}}{\sqrt{1 + (\bar{\sigma}_{ij}/\sigma_o)^2}} + t_{ij}\right) \\ T_{ij} &\sim H\left(0, \sigma_o^2\right).\end{aligned} \tag{A3}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

It follows that the expectation of $Y_{ij}|\boldsymbol{x}_{ij}$ is given by

$$\mu_{ij} = h\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right). \tag{A4}$$

□

Appendix A.2. Proof of Proposition 2

To prove Proposition 2, we make use of the following two Lemmas.

Lemma A1. *Consider the function $\varphi_1 : \mathbb{R}^p \to \mathbb{R}^p$ defined for $\mathbf{z} \in \mathbb{R}^p$ as*

$$\varphi_1(\mathbf{z}) = \frac{\mathbf{z}}{\sqrt{1+(\bar{\sigma}/\sigma_o)^2}} \tag{A5}$$

where $\bar{\sigma}^2 = \mathbf{z}^\top \boldsymbol{S}\mathbf{z}$, $\boldsymbol{S} \in \mathbb{R}^{p\times p}$ is a semi-positive definite matrix, and $\sigma_o > 0$ is a positive constant. The function φ_1 is injective, and accordingly, for any $\mathbf{a} \in \mathbb{R}^p$, the equation $\varphi_1(\mathbf{z}) = \varphi_1(\mathbf{a})$ has a unique solution $\mathbf{z} = \mathbf{a}$.

Proof. Denoting $\boldsymbol{I}_p$ the $p \times p$ identity matrix, and setting $v = \sqrt{1+(\bar{\sigma}/\sigma_o)^2}$, the Jacobian matrix $\dot{\varphi}_1(\mathbf{z}) = \frac{\partial \varphi_1(\mathbf{z})}{\partial \mathbf{z}^\top}$ of $\varphi_1(\mathbf{z})$ is given by $\dot{\varphi}_1(\mathbf{z}) = \frac{1}{v}\boldsymbol{I}_p + \mathbf{z}\frac{\partial}{\partial \mathbf{z}^\top}\left(\frac{1}{v}\right)$. Further calculus yields

$$\dot{\varphi}_1(\mathbf{z}) = \frac{1}{v}\left[\boldsymbol{I}_p - \frac{1}{\sigma_o^2 v^2}\mathbf{z}\mathbf{z}^\top \boldsymbol{S}\right].$$

By the matrix determinant lemma [103], the determinant of $\dot{\varphi}_1$ is given by

$$\begin{aligned}|\dot{\varphi}_1(\mathbf{z})| &= v^{-p}\left(1 - \frac{1}{\sigma_o^2}\frac{\mathbf{z}^\top \boldsymbol{S}\mathbf{z}}{\left(1+\frac{\mathbf{z}^\top \boldsymbol{S}\mathbf{z}}{\sigma_o^2}\right)}\right) \\ &= v^{-p}\left(1+\frac{\mathbf{z}^\top \boldsymbol{S}\mathbf{z}}{\sigma_o^2}\right)^{-1} = v^{-p-2}.\end{aligned}$$

Since $\boldsymbol{S}$ is semi-positive definite, $\mathbf{z}^\top \boldsymbol{S}\mathbf{z} \geq 0$ and $v > 0$ for any $\mathbf{z} \in \mathbb{R}^p$. It follows that $|\dot{\varphi}_1(\mathbf{z})| \neq 0$ for all $\mathbf{z} \in \mathbb{R}^p$. □

Lemma A2. *Consider the function $\varphi_2 : \mathbb{R}^p \to \mathbb{R}^p$ defined for $\mathbf{z} \in \mathbb{R}^p$ and $\mathbf{z} \neq \mathbf{0}$ as*

$$\varphi_2(\mathbf{z}) = \left(\mathbf{z}_0^\top \mathbf{z}\right)\cdot \mathbf{z} \tag{A6}$$

where $\mathbf{z}_0 \in \mathbb{R}^p$ with linearly independent components. The function φ_2 is injective, and accordingly, for any $\mathbf{z} \in \mathbb{R}^p$ and $\mathbf{z} \neq \mathbf{0}$, the equation $\varphi_2(\mathbf{a}) = \varphi_2(\mathbf{z})$ has a unique solution $\mathbf{a} = \mathbf{z}$.

Proof. The Jacobian matrix of $\varphi_2(\mathbf{z})$ is given by $\dot{\varphi}_2(\mathbf{z}) = \left(\mathbf{z}_0^\top \mathbf{z}\right)\cdot\left(\boldsymbol{I}_p + \frac{\mathbf{z}\mathbf{z}_0^\top}{\mathbf{z}_0^\top \mathbf{z}}\right)$. Note that the second term in the square bracket is well defined since $\mathbf{z}_0$ have linearly independent components and $\mathbf{z} \neq \mathbf{0}$ imply that $\mathbf{z}_0^\top \mathbf{z} \neq 0$. The determinant of $\dot{\varphi}_2$ is thus

$$|\dot{\varphi}_2(\mathbf{z})| = \left(\mathbf{z}_0^\top \mathbf{z}\right)^p\left[1+\frac{\mathbf{z}_0^\top \mathbf{z}}{\mathbf{z}_0^\top \mathbf{z}}\right] = 2\left(\mathbf{z}_0^\top \mathbf{z}\right)^p.$$

It follows that $|\dot{\varphi}_2(\mathbf{z})| \neq 0$ for all non null $\mathbf{z} \in \mathbb{R}^p$. □

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Proof of Proposition 2. Let us denote $(\Omega, \mathcal{A}, \mathbb{P})$ the probability space on which the random vector $(Y, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_q, \boldsymbol{U}, \boldsymbol{V})$ is defined. Suppose that $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}^*)$ almost surely. Then assumptions **C**1 and **C**2 imply the existence of a positive constant c_1 such that for every $\mathbf{x}_j \in \mathbb{C}_j$, $\mathbf{u} \in \mathbb{C}_u$, $\mathbf{v} \in \mathbb{C}_v$, and $\boldsymbol{\theta} \in \mathcal{B}_p \times \mathcal{B}_d \times \mathcal{B}_r$, $c_1 < \mathbb{P}(Y = 1|\mathbf{x}_1, \cdots, \mathbf{x}_q, \mathbf{u}, \mathbf{v}) < 1 - c_1$. It follows that there exists $\omega \in \Omega$, outside the region (of measure zero) where $\ell(\boldsymbol{\theta}) \neq \ell(\boldsymbol{\theta}^*)$ such that $Y(\omega) = 1$ given that $\mathbf{X}_j = \boldsymbol{x}_j$, $\mathbf{U} = \boldsymbol{u}$ and $\mathbf{V} = \boldsymbol{v}$. For such a ω, the equality $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}^*)$ reads $\mu = \mu^*$.

For simplicity in the presentation we define $\varrho = \lambda \prod_{j=1}^{q} \mu_j$ so that μ can be rewritten as $\mu = \alpha\lambda + (1-\alpha)\varrho$. We first show that the quantities α, λ and ϱ are individually identifiable. In case either α and λ are known constants in the MSB model, we only need ϱ to be identifiable. Otherwise, from $\alpha = h(\boldsymbol{u}^\top \boldsymbol{\gamma})$ or $\lambda = h(\boldsymbol{v}^\top \boldsymbol{\delta})$, the identifiability of α or λ is reduced to the identifiability problem in the standard binomial regression model. Specifically, assumptions **C**1 and **C**2 ensure that $\alpha = \alpha^*$ and $\lambda = \lambda^*$ respectively imply that $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ (see e.g. [104]). As for ϱ, $\varrho = \varrho^*$ reads

$$\lambda \prod_{j=1}^{q} h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{\sqrt{1 + (\bar{\sigma}_j/\sigma_o)^2}}\right) = \lambda^* \prod_{j=1}^{q} h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{\sqrt{1 + (\bar{\sigma}_j^*/\sigma_o)^2}}\right)$$

where $\bar{\sigma}_j = \sqrt{\boldsymbol{\beta}_j^\top \boldsymbol{S} \boldsymbol{\beta}_j}$, $\bar{\sigma}_j^* = \sqrt{\boldsymbol{\beta}_j^{*\top} \boldsymbol{S} \boldsymbol{\beta}_j^*}$. Set $v_j = \sqrt{1 + (\bar{\sigma}_j/\sigma_o)^2}$ and $v_j^* = \sqrt{1 + (\bar{\sigma}_j^*/\sigma_o)^2}$. By the positivity of all h values under condition **C**2, and given that $\lambda = \lambda^*$, we can write for any $j = 1, 2, \cdots, q$:

$$\frac{h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right)}{h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)} = \frac{\prod_{k \neq j} h\left(\frac{\boldsymbol{x}_k^\top \boldsymbol{\beta}_k^*}{v_k^*}\right)}{\prod_{k \neq j} h\left(\frac{\boldsymbol{x}_k^\top \boldsymbol{\beta}_k}{v_k}\right)}. \tag{A7}$$

Differentiating with respect to T_j, a continuous predictor in $\boldsymbol{x}_j$ with regression coefficient $\beta_{jt} \neq 0$ (or $\beta_{jt}^* \neq 0$) under condition **C**3, we get

$$\frac{\frac{\beta_{jt}}{v_j} \dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right) h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right) - \frac{\beta_{jt}^*}{v_j^*} h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right) \dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)}{\left[h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)\right]^2} = 0 \tag{A8}$$

where $\dot{h}(x) = \frac{\partial h(x)}{\partial x}$ (density function corresponding to the link h). The right hand side of the result is zero because the right hand side of (A7) does not depend on $\boldsymbol{x}_j$. The equality (A8) implies

$$\frac{\beta_{jt} v_j^*}{\beta_{jt}^* v_j} = \frac{h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right) \dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)}{\dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right) h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)}. \tag{A9}$$

Further differentiation of both sides of Equation (A9) with respect to T_j gives

$$0 \cdot \frac{\dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)}{\dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right)} + \frac{h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right)}{h\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right)} \frac{\frac{\beta_{jt}^*}{v_j^*} \ddot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right) \dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right) - \frac{\beta_{jt}}{v_j} \dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j^*}{v_j^*}\right) \ddot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right)}{\left[\dot{h}\left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}_j}{v_j}\right)\right]^2} = 0$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

where the factor zero in the first term on the left hand side results from (A8), and the right hand side is zero because the left hand side of (A9) does not depend on $\boldsymbol{x}_j$. The last equality implies

$$\frac{\beta_{jt}v_j^*}{\beta_{jt}^*v_j} = \frac{\dot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)\ddot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}{\ddot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)\dot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}. \tag{A10}$$

Equations (A9) and (A10) gives the double equality

$$\frac{\beta_{jt}v_j^*}{\beta_{jt}^*v_j} = \frac{h\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)\dot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}{\dot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)h\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)} = \frac{\dot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)\ddot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}{\ddot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)\dot{h}\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}.$$

which is central to our proof of $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^*$. We next consider specific cases of link functions.

Case of logit link: $h(x) = (1+e^{-x})^{-1}$
Here, $\dot{h}(x) = h(x)[1-h(x)]$, hence $\frac{h(x)}{\dot{h}(x)} = \frac{1}{1-h(x)} = 1+e^x$. Equation (A9) becomes

$$\frac{\beta_{jt}v_j^*}{\beta_{jt}^*v_j} = \frac{1+\exp\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)}{1+\exp\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}.$$

Differentiating both sides of the last equality with respect to T_j yields $\left(\frac{\boldsymbol{\beta}_j}{v_j} - \frac{\boldsymbol{\beta}_j^*}{v_j^*}\right)^\top \boldsymbol{x}_j = 0$. Along with the condition **C**1, this implies that $\frac{\boldsymbol{\beta}_j}{v_j} = \frac{\boldsymbol{\beta}_j^*}{v_j^*}$, and by Lemma A1, we get $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^*$.

Case of probit link: $h(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$
We have $\dot{h}(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, and $\ddot{h}(x) = -x\dot{h}(x)$, hence $\frac{\ddot{h}(x)}{\dot{h}(x)} = -x$. Note that by the condition **C**1 and the assumption $\beta_{jt} \neq 0$, we have $\boldsymbol{x}_j^\top\boldsymbol{\beta}_j \neq 0$. Equation (A10) thus reads

$$\frac{\beta_{jt}v_j^*}{\beta_{jt}^*v_j} = \frac{\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}\right)}{\left(\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j}\right)}, \quad \text{which is rearranged as} \quad \frac{\beta_{jt}}{v_j}\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j}{v_j} = \frac{\beta_{jt}^*}{v_j^*}\frac{\boldsymbol{x}_j^\top\boldsymbol{\beta}_j^*}{v_j^*}.$$

Then, because the assumption $\beta_{jt} \neq 0$ gives $\boldsymbol{\beta}_j \neq \mathbf{0}$ and condition **C**1 ensures $\boldsymbol{\beta}_j^\top\boldsymbol{x}_j \neq 0$, by Lemma A2, we have $\frac{\boldsymbol{\beta}_j}{v_j} = \frac{\boldsymbol{\beta}_j^*}{v_j^*}$, which then gives $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^*$ by Lemma A1. We have overall shown that under the logit or probit link function, $\varrho = \varrho^*$ implies that $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^*$ for $j = 1, 2, \cdots, q$.

It remains to show that $\mu = \mu^*$ implies that $\alpha = \alpha^*$, $\lambda = \lambda^*$ and $\varrho = \varrho^*$. In case both α and λ are known constants in the MSB model, $\mu = \alpha\lambda + (1-\alpha)\varrho$ implies that μ as a function of ϱ is invertible ($\varrho = \frac{\mu-\alpha\lambda}{(1-\alpha)}$ given that $\alpha \in (0,1)$ and $\mu \geq \alpha\lambda$). Hence $\mu = \mu^*$ implies $\varrho = \varrho^*$. If both α and λ are not known constants, we have for all $j = 1, 2, \cdots, q$, $\frac{\partial}{\partial T_j}\frac{\varrho_0}{\varrho_0^*} = 0$ (where $\varrho_0 = \prod_{j=1}^q \mu_j$) which leads to $\varrho_0 = \varrho_0^*$ by the above argument used for $\boldsymbol{\beta}_j$. In case exactly one of α and λ is a known constant (say α is a known constant), the second (say λ) can be uniquely recovered from knowing μ and ϱ_0.

If none of α and λ is a known constant, then the condition **C**3 ensures that there is a continuous predictor T_α in the linear predictor of α or T_λ in the linear predictor of λ. Then, $\mu = \mu^*$ reads $\alpha\lambda + (1-\alpha)\lambda\varrho_0 = \alpha^*\lambda^* + (1-\alpha^*)\lambda^*\varrho_0^*$. Suppose that we have T_α, then we get $\frac{\partial}{\partial T_\lambda}\frac{\lambda}{\lambda^*} = 0$ and $\delta = \delta^*$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

follows. This results in $\lambda = \lambda^*$. Since $\varrho_0 = \varrho_0^*$ and $\lambda = \lambda^*$, it follows that $\alpha = \alpha^*$. If we have T_λ, then we get $\frac{\partial}{\partial T_\alpha}\frac{\lambda}{\lambda^*} = 0$ giving $\alpha = \alpha^*$. From $\varrho_0 = \varrho_0^*$ and $\alpha = \alpha^*$, it follows that $\lambda = \lambda^*$. □

Appendix B. Score Vector and Information Matrix

Appendix B.1. Maximum Likelihood

The Fisher (expected) information matrix required in the penalized log-likelihood function (15) is defined as the variance-covariance matrix of the score vector $\dot{\boldsymbol{\ell}}_n = \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}}$ [83]:

$$\boldsymbol{\mathcal{E}}_n(\boldsymbol{\theta}) = Var\Big[\dot{\boldsymbol{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y})\Big]. \tag{A11}$$

From (14), the score vector is given by

$$\dot{\boldsymbol{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i - n_i\mu_i}{\mu_i(1-\mu_i)}\dot{\boldsymbol{\mu}}_i \tag{A12}$$

where $\dot{\boldsymbol{\mu}}_i = \frac{\partial \mu_i}{\partial \boldsymbol{\theta}}$ is the vector of first order partial derivatives of μ_i with respect to the elements of the parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_q^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top\right)^\top$. From (12), $\dot{\boldsymbol{\mu}}_i$ is given by:

$$\dot{\boldsymbol{\mu}}_i = \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}_1^\top}, \cdots, \frac{\partial \mu_i}{\partial \boldsymbol{\beta}_q^\top}, \frac{\partial \mu_i}{\partial \boldsymbol{\gamma}^\top}, \frac{\partial \mu_i}{\partial \boldsymbol{\delta}^\top}\right)^\top \tag{A13}$$

where

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}_j} = \lambda_i(1-\alpha_i)\left[\prod_{k\neq j} h\left(\frac{\eta_{ik}}{\sqrt{1+(\bar{\sigma}_{ik}/\sigma_o)^2}}\right)\right]\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j} \quad (\text{for} \quad j = 1, \cdots, q), \tag{A14a}$$

$$\frac{\partial \mu_i}{\partial \boldsymbol{\gamma}} = \lambda_i\left[1 - \prod_{j=1}^{q} h\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\right]\dot{h}\left(\boldsymbol{u}_i^\top\boldsymbol{\gamma}\right)\boldsymbol{u}_i, \tag{A14b}$$

$$\frac{\partial \mu_i}{\partial \boldsymbol{\delta}} = \left[\alpha_i + (1-\alpha_i)\prod_{j=1}^{q} h\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\right]\dot{h}\left(\boldsymbol{v}_i^\top\boldsymbol{\delta}\right)\boldsymbol{v}_i \tag{A14c}$$

with $\dot{h}(x) = \frac{\partial h(x)}{\partial x}$ (density function corresponding to the link h),

$$\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j} = \dot{h}\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\boldsymbol{\tau}_{ij}, \quad \text{and} \tag{A14d}$$

$$\boldsymbol{\tau}_{ij} = \frac{1}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\left[\boldsymbol{x}_{ij} - \frac{\eta_{ij}}{\sigma_o^2+\bar{\sigma}_{ij}^2}\boldsymbol{S}_{ij}\boldsymbol{\beta}_j\right]. \tag{A14e}$$

Note that in the absence of measurement errors ($\boldsymbol{S}_{ij} = \boldsymbol{0}$), $\boldsymbol{\tau}_{ij} = \boldsymbol{x}_{ij}$ so that $\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j} = \dot{h}(\eta_{ij})\boldsymbol{x}_{ij}$. Since $\dot{\boldsymbol{\mu}}_i$ does not depend on y_i, and $Var\left[\frac{y_i - n_i\mu_i}{\mu_i(1-\mu_i)}\right] = \frac{Var[y_i]}{\mu_i^2(1-\mu_i)^2} = \frac{n_i}{\mu_i(1-\mu_i)}$, we obtain the Fisher information matrix

$$\boldsymbol{\mathcal{E}}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{n_i}{\mu_i(1-\mu_i)}\dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_i^\top. \tag{A15}$$

When the ML estimator $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\gamma}}^\top, \widehat{\boldsymbol{\delta}}^\top\right)^\top$ of $\boldsymbol{\theta}$ exists, large sample inference on $\widehat{\boldsymbol{\theta}}$ relies on the asymptotic covariance matrix as usually in GLMs. A candidate for the asymptotic covariance matrix is

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

the inverse of the observed information matrix (evaluated at $\widehat{\boldsymbol{\theta}}$) [105], given for the MSB model (8) by $\boldsymbol{\Sigma}_n = \left[\mathcal{I}_n(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1}$ where

$$\mathcal{I}_n(\boldsymbol{\theta}|\boldsymbol{y}) = -\ddot{\ell}_n(\boldsymbol{\theta}|\boldsymbol{y})$$

with $\ddot{\boldsymbol{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \dot{\boldsymbol{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}^\top}$. By differentiating (A12) with respect to $\boldsymbol{\theta}^\top$, we obtain

$$\mathcal{I}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n}\left[\left(\frac{y_i}{\mu_i^2} + \frac{n_i - y_i}{(1-\mu_i)^2}\right)\dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_i^\top - \frac{y_i - n_i\mu_i}{\mu_i(1-\mu_i)}\ddot{\boldsymbol{\mu}}_i\right] \tag{A16}$$

where $\ddot{\boldsymbol{\mu}}_i = \frac{\partial^2 \mu_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ is obtained from (A13) as

$$\ddot{\boldsymbol{\mu}}_i = \begin{bmatrix} \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_2^\top} & \cdots & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_q^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\gamma}^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\delta}^\top} \\ \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_1^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2^\top} & \cdots & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_q^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\gamma}^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\delta}^\top} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_q\partial\boldsymbol{\beta}_1^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_q\partial\boldsymbol{\beta}_2^\top} & \cdots & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_q\partial\boldsymbol{\beta}_q^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_q\partial\boldsymbol{\gamma}^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\beta}_q\partial\boldsymbol{\delta}^\top} \\ \frac{\partial^2\mu_i}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}_1^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}_2^\top} & \cdots & \frac{\partial^2\mu_i}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}_q^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\delta}^\top} \\ \frac{\partial^2\mu_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\beta}_1^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\beta}_2^\top} & \cdots & \frac{\partial^2\mu_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\beta}_q^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\gamma}^\top} & \frac{\partial^2\mu_i}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top} \end{bmatrix} \tag{A17}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

where for $j = 1, \cdots, q$ and on setting $\ddot{h}(x) = \frac{\partial^2 h(x)}{\partial x^2}$:

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\beta}_j \boldsymbol{\beta}_j^\top} = \lambda_i(1-\alpha_i)\left[\prod_{k\neq j} h\left(\frac{\eta_{ik}}{\sqrt{1+(\bar{\sigma}_{ik}/\sigma_o)^2}}\right)\right]\frac{\partial^2 \mu_{ij}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top}, \tag{A18a}$$

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top} = \lambda_i\left[1-\prod_{j=1}^{q} h\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\right]\ddot{h}\left(\boldsymbol{u}_i^\top \boldsymbol{\gamma}\right)\boldsymbol{u}_i \boldsymbol{u}_i^\top, \tag{A18b}$$

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} = \left[\alpha_i + (1-\alpha_i)\prod_{j=1}^{q} h\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\right]\ddot{h}\left(\boldsymbol{v}_i^\top \boldsymbol{\delta}\right)\boldsymbol{v}_i \boldsymbol{v}_i^\top, \tag{A18c}$$

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k^\top} = \lambda_i(1-\alpha_i)\left[\prod_{l\neq j,k} h\left(\frac{\eta_{il}}{\sqrt{1+(\bar{\sigma}_{il}/\sigma_o)^2}}\right)\right]\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j}\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_k^\top} \quad (\text{for} \quad k \neq j), \tag{A18d}$$

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\gamma}^\top} = -\lambda_i\left[\prod_{k\neq j} h\left(\frac{\eta_{ik}}{\sqrt{1+(\bar{\sigma}_{ik}/\sigma_o)^2}}\right)\right]\dot{h}\left(\boldsymbol{u}_i^\top \boldsymbol{\gamma}\right)\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j}\boldsymbol{u}_i^\top, \tag{A18e}$$

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\delta}^\top} = (1-\alpha_i)\left[\prod_{k\neq j} h\left(\frac{\eta_{ik}}{\sqrt{1+(\bar{\sigma}_{ik}/\sigma_o)^2}}\right)\right]\dot{h}\left(\boldsymbol{v}_i^\top \boldsymbol{\delta}\right)\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j}\boldsymbol{v}_i^\top, \tag{A18f}$$

$$\frac{\partial^2 \mu_i}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\delta}^\top} = \left[1-\prod_{j=1}^{q} h\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\right]\dot{h}\left(\boldsymbol{u}_i^\top \boldsymbol{\gamma}\right)\dot{h}\left(\boldsymbol{v}_i^\top \boldsymbol{\delta}\right)\boldsymbol{u}_i \boldsymbol{v}_i^\top, \tag{A18g}$$

$$\frac{\partial^2 \mu_{ij}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} = \ddot{h}\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\boldsymbol{\tau}_{ij}\boldsymbol{\tau}_{ij}^\top + \dot{h}\left(\frac{\eta_{ij}}{\sqrt{1+(\bar{\sigma}_{ij}/\sigma_o)^2}}\right)\dot{\boldsymbol{\tau}}_{ij}, \tag{A18h}$$

$$\text{with} \quad \dot{\boldsymbol{\tau}}_{ij} = \frac{\sigma_o}{\left(\sigma_o^2+\bar{\sigma}_{ij}^2\right)^{3/2}}\left[\frac{3\eta_{ij}}{\sigma_o^2+\bar{\sigma}_{ij}^2}\boldsymbol{S}_{ij}\boldsymbol{\beta}_j\boldsymbol{\beta}_j^\top\boldsymbol{S}_{ij} - \eta_{ij}\boldsymbol{S}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}_j^\top\boldsymbol{S}_{ij} - \boldsymbol{S}_{ij}\boldsymbol{\beta}_j\boldsymbol{x}_{ij}^\top\right]. \tag{A18i}$$

An alternative to $\mathcal{I}_n(\boldsymbol{\theta}|\boldsymbol{y})$ used in $\boldsymbol{\Sigma}_n = \left[\mathcal{I}_n(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1}$ is the Fisher information matrix $\boldsymbol{\mathcal{E}}_n(\boldsymbol{\theta})$. Using $\boldsymbol{\mathcal{E}}_n$, we have the potential asymptotic covariance matrix $\boldsymbol{\Gamma}_n = \left[\boldsymbol{\mathcal{E}}_n(\widehat{\boldsymbol{\theta}})\right]^{-1}$ as an alternative to $\boldsymbol{\Sigma}_n$. For the purpose of constructing asymptotic confidence intervals for estimates, recent studies suggest that the Fisher information $\boldsymbol{\mathcal{E}}_n(\widehat{\boldsymbol{\theta}})$ performs at least as well as the observed information matrix $\boldsymbol{\mathcal{I}}_n(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})$ [106–108]. Interestingly, $\boldsymbol{\mathcal{E}}_n$ is also less expensive to compute as compared to $\boldsymbol{\mathcal{I}}_n$. Indeed, Equation (A17) shows that even after computing the score (A12), a lot of additional computations are required to obtain $\mathcal{I}_n(\boldsymbol{\theta}|\boldsymbol{y})$. However, $\mathcal{I}_n(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})$ is usually available as a by-product from optimization routines in popular statistical softwares (see e.g. *optim* in R [84]).

Appendix B.2. Penalized Maximum Likelihood

For PML estimation based on $\tilde{\ell}_n$ (15), the penalized score $\dot{\tilde{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \tilde{\ell}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is given by

$$\dot{\tilde{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \dot{\ell}_n(\boldsymbol{\theta}|\boldsymbol{y}) + \frac{1}{2}\frac{\partial log|\mathcal{E}_n(\boldsymbol{\theta})|}{\partial \boldsymbol{\theta}} \tag{A19}$$

where $\dot{\ell}_n$ is the score given in (A12), and by the chain rule, the double score penalty satisfies

$$\frac{\partial log|\mathcal{E}_n(\boldsymbol{\theta})|}{\partial \boldsymbol{\theta}_j} = \text{Tr}\left[\left[\mathcal{E}_n(\boldsymbol{\theta})\right]^{-1}\frac{\partial \mathcal{E}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right]. \tag{A20}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

The required derivatives of $\mathcal{E}_n(\boldsymbol{\theta})$ are obtained from $\frac{\partial \mathcal{E}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\frac{n_i}{\mu_i(1-\mu_i)}\dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_i^\top\right)$ as

$$\frac{\partial \mathcal{E}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{n}\left(\frac{n_i(2\mu_i-1)}{\mu_i^2(1-\mu_i)^2}\dot{\boldsymbol{\mu}}_{i,j}\right)\dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_i^\top + \frac{n_i}{\mu_i(1-\mu_i)}\boldsymbol{M}(i,j),$$

where $\dot{\boldsymbol{\mu}}_{i,j}$ is the jth element of $\dot{\boldsymbol{\mu}}_i$ given by Equation (A13), and the matrix $\boldsymbol{M}(i,j) = \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_i^\top\right)$ has u,vth element $\boldsymbol{M}(i,j)_{uv} = \frac{\partial}{\partial \boldsymbol{\theta}_j}(\dot{\boldsymbol{\mu}}_{i,u}\dot{\boldsymbol{\mu}}_{i,v})$, i.e.,

$$\boldsymbol{M}(i,j) = \ddot{\boldsymbol{\mu}}_{i,\cdot j}\dot{\boldsymbol{\mu}}_i^\top + \dot{\boldsymbol{\mu}}_i\ddot{\boldsymbol{\mu}}_{i,\cdot j}^\top,$$

with $\ddot{\boldsymbol{\mu}}_{i,\cdot j}$ the jth column of the hessian matrix $\ddot{\boldsymbol{\mu}}_i$ given in Equation (A17). Note that Equation (A20) can be rewritten as

$$\frac{\partial log|\mathcal{E}_n(\boldsymbol{\theta})|}{\partial \boldsymbol{\theta}_j} = \dot{\mathcal{E}}_{n,j}^\top(\boldsymbol{\theta})\ \mathrm{vec}\Big([\mathcal{E}_n(\boldsymbol{\theta})]^{-1}\Big)$$

where $\dot{\mathcal{E}}_{n,j}(\boldsymbol{\theta}) = \mathrm{vec}\left(\frac{\partial \mathcal{E}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right) = \frac{\partial \mathrm{vec}(\mathcal{E}_n(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_j}$ and vec is the usual operator which stacks the columns of its matrix argument. Denoting $\bigotimes$ the Kronecker (direct) product, and applying Equation (2) in [109], we get

$$\dot{\mathcal{E}}_{n,j}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left(\frac{n_i(2\mu_i-1)}{\mu_i^2(1-\mu_i)^2}\dot{\boldsymbol{\mu}}_{i,j}\right)\dot{\boldsymbol{\mu}}_i \bigotimes \dot{\boldsymbol{\mu}}_i + \frac{n_i}{\mu_i(1-\mu_i)}\left[\dot{\boldsymbol{\mu}}_i \bigotimes \ddot{\boldsymbol{\mu}}_{i,\cdot j} + \ddot{\boldsymbol{\mu}}_{i,\cdot j} \bigotimes \dot{\boldsymbol{\mu}}_i\right].$$

Further using Equation (1) in [109] yields

$$\dot{\mathcal{E}}_{n,j}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left(\frac{n_i(2\mu_i-1)}{\mu_i^2(1-\mu_i)^2}\dot{\boldsymbol{\mu}}_{i,j}\right)\dot{\boldsymbol{\mu}}_i \bigotimes \dot{\boldsymbol{\mu}}_i + \frac{n_i}{\mu_i(1-\mu_i)}(\boldsymbol{K}_{m,m} + \boldsymbol{I}_{m^2})\left(\ddot{\boldsymbol{\mu}}_{i,\cdot j} \bigotimes \dot{\boldsymbol{\mu}}_i\right)$$

where $\boldsymbol{K}_{m,m}$ is the commutation matrix [110] and m denotes the length of the parameter vector $\boldsymbol{\theta}$. Collecting the column vectors $\dot{\mathcal{E}}_{n,j}$, we obtain

$$\frac{\partial log|\mathcal{E}_n(\boldsymbol{\theta})|}{\partial \boldsymbol{\theta}} = \dot{\mathcal{E}}_n^\top(\boldsymbol{\theta})\ \mathrm{vec}\Big([\mathcal{E}_n(\boldsymbol{\theta})]^{-1}\Big) \tag{A21a}$$

where $\dot{\mathcal{E}}_n$ is a $m^2 \times m$ jerk matrix defined as $\dot{\mathcal{E}}_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}^\top}\mathrm{vec}(\mathcal{E}_n(\boldsymbol{\theta}))$ and given by

$$\dot{\mathcal{E}}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}\frac{n_i}{\mu_i(1-\mu_i)}\left[\frac{2\mu_i-1}{\mu_i(1-\mu_i)}\left(\dot{\boldsymbol{\mu}}_i \bigotimes \dot{\boldsymbol{\mu}}_i\right)\dot{\boldsymbol{\mu}}_i^\top + (\boldsymbol{K}_{m,m} + \boldsymbol{I}_{m^2})\left(\ddot{\boldsymbol{\mu}}_i \bigotimes \dot{\boldsymbol{\mu}}_i\right)\right]. \tag{A21b}$$

Solving the penalized score equation $\dot{\tilde{\ell}}_n(\boldsymbol{\theta}|\boldsymbol{y}) = \mathbf{0}$ (with $\dot{\tilde{\ell}}_n$ given by Equation (A19)) produces the PML estimates $\widetilde{\boldsymbol{\theta}} = \left(\widetilde{\boldsymbol{\beta}}^\top, \widetilde{\boldsymbol{\gamma}}^\top, \widetilde{\boldsymbol{\delta}}^\top\right)^\top$ of $\boldsymbol{\theta}$. For large sample inference on $\widetilde{\boldsymbol{\theta}}$, the Fisher information based asymptotic covariance matrix is given by $\boldsymbol{\Gamma}_n = \left[\boldsymbol{\mathcal{E}}_n(\widetilde{\boldsymbol{\theta}})\right]^{-1}$. Indeed, the penalized Fisher information matrix (covariance matrix of $\dot{\tilde{\ell}}_n$) is, by Equation (A11) (the same as for $\dot{\ell}_n$), given by Equation (A15) because $\dot{\tilde{\ell}}_n$ is equal to $\dot{\ell}_n$ up to an additive constant (independent of the response Y) which does not affect covariance matrix calculation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

Appendix C. Simulation Details

Appendix C.1. Simulation Design

We consider the following extension of the MSB data generation process considered in section 3:

$$Y_i \overset{ind}{\sim} \mathcal{B}in(1, \mu_i), \tag{A22a}$$

$$\text{where} \qquad \mu_i = \lambda \prod_{j=1}^{q} h(\eta_{ij}), \tag{A22b}$$

$$\eta_{ij} = \beta_0 + \beta_1 X_{ij}. \tag{A22c}$$

In (A22), Y_i is the binary response for unit i, $\lambda \in (0, 1]$ is an upper limit for the success probability of the binary response Y, h is the logit link function defined in (3), X denotes the explanatory variable following a predictor distribution denoted PD, and β_0 and β_1 are intercept and slopes.

Note that if $\lambda = 1$ with $q = 1$, then (A22) is a standard simple logistic data generation process. For $\lambda < 1$ and $q = 1$, (A22) represents a multistage process where only one stage is observed. We varied the upper limit λ of the success probability of the response Y from 0.5 to 1, varying regression coefficients as $\beta_0 = 3, 8$ and $\beta_1 = -3, -2, -1, 0$, and the number n of units from 100 to 2000.

Under model (A22), the true success probability curves are shown in Figure A1 for $\beta_0 = 3$, $\beta_1 = -1$ and selected λ values from 0.1 to 1. For a predictor X, we first considered a uniform distribution. For instance, with $\beta_0 = 3$ and $\beta_1 = -1$, we simulated X values uniformly from the interval $[-2, 8]$ (Figure A2), having expectation $\mu_x = 3$ and standard deviation $\sigma_x = 2.89$, and spanning 99% probability mass of the standard logistic distribution centered at 3.

It is well established that the distribution of the explanatory variable has a large impact on inference in logistic regression [111]. To investigate this effect in our MSB framework, we considered two alternative continuous probability distributions for X, including the normal (symmetric) and the lognormal (asymmetric) distributions. Normal covariate values were simulated with mean $\mu_w = 3$ and standard deviation $\sigma_x = 1.6$. We used a shifted lognormal distribution, obtained by subtracting two from log normal variates with mean $\mu_x = 2.65$ and standard deviation $\sigma_x = 1.41$ (log scale mean and standard are 1.1 and 0.5, respectively). For both normal and lognormal covariates, the chosen distribution parameters ensure that X values cover the interval $[-2, 8]$ with probability 99% when $\beta_0 = 3$ and $\beta_1 = -1$ (Figure A2). The lognormal distribution giving more weight to X values along the asymptote of the success probability curve ($X \in [-0.89, 2.91]$; Figure A2) is expected to result in more precise (lower variance) estimates of the asymptote λ (irrespective of the true λ value), but less accurate (more biased/less precise) estimates of β_0 and β_1 when the true λ value is low.

A summary of our simulation factors, including λ, β_0, β_1, PD, and sample size (n) is shown in Table A1. As performance measures, we computed the relative bias and root mean square error ($rmse$) of the estimates of model parameters λ, β_0 and β_1, and the $mad(\hat{\mu})$ and $R^2(\hat{\mu})$ of predicted success probabilities as defined in Section 3.1, based on 2000 independent replicates of each simulation setting.

Table A1. Levels of simulation factors

Simulation factor	Notation	Levels
Sample size	n	100, 200, 300, 500, 750, 1000, 1500, 2000
Upper limit of success probability	λ	0.5, 0.75, 0.9, 1
Intercept	β_0	3, 8
Slope	β_1	-3, -2, -1
Distribution of the predictor	PD	Uniform, Normal, Log-normal

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

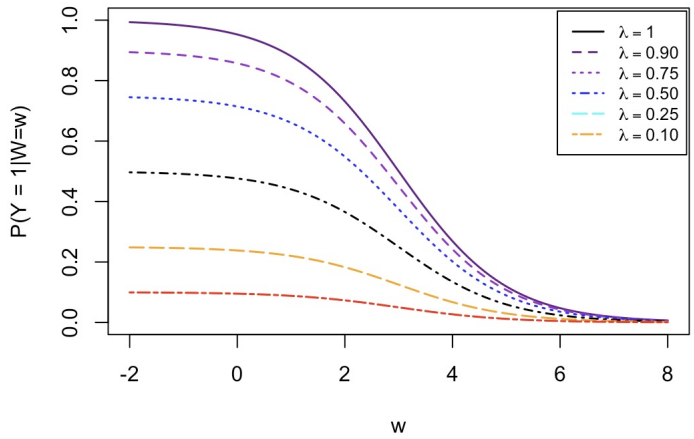


Figure A1. Success probability from model (A22) as a function the explanatory variable X for selected asymptote values λ and $\beta_0 = 3$ and $\beta_1 = -1$.

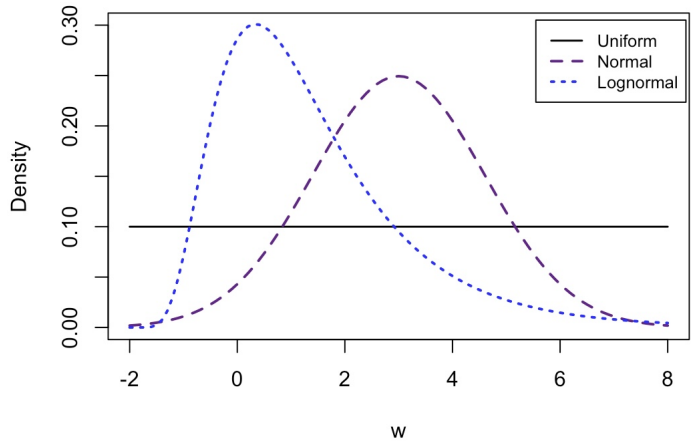


Figure A2. Probability distributions considered for the covariate X in model (A22) with $\beta_0 = 3$ and $\beta_1 = -1$. All density curves cover the region $w \in [-2, 8]$ with probability 99% or more (100% for the uniform distribution). The normal density is above the uniform density curve over the region $X \in [0.84, 5.16]$ where the normal has 90% probability against 43% for the uniform. The lognormal density is above the uniform density curve over the region $X \in [-0.89, 2.91]$ where the lognormal has 81% probability against 38% for the uniform. For general β_0 and β_1 values, X is rescaled (linear transformation) so as to preserve the logit scale range of the linear predictor $\eta = \beta_0 + \beta_1 X$.

Appendix C.2. Additional Simulation Results

Appendix C.2.1. Bias

Figures A3–A14 present the relative bias in the estimates $\hat{\lambda}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ for model parameters $\lambda \in [0.25, 1]$, $\beta_0 \in \{3, 8\}$ and $\beta_1 \in \{-3, -2, -1\}$ and sample size $n \in [100, 2000]$. The relative bias generally approaches zero as n increases. For $\hat{\lambda}$, the bias is generally negative for high true λ values (close to one), and positive for low λ values ($0.25 \leq \lambda < 0.5$).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

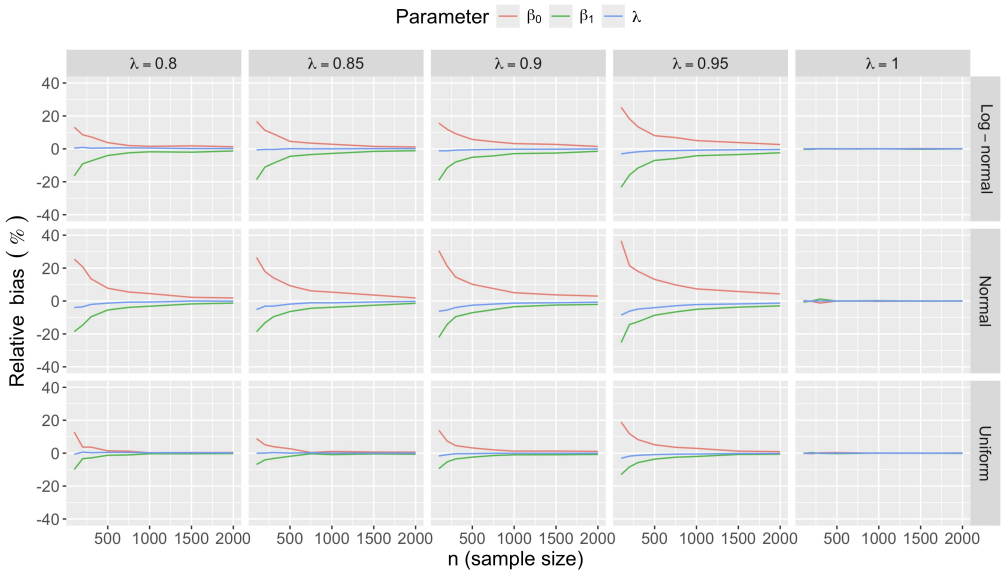


Figure A3. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -1$, and $\lambda \geq 0.8$.

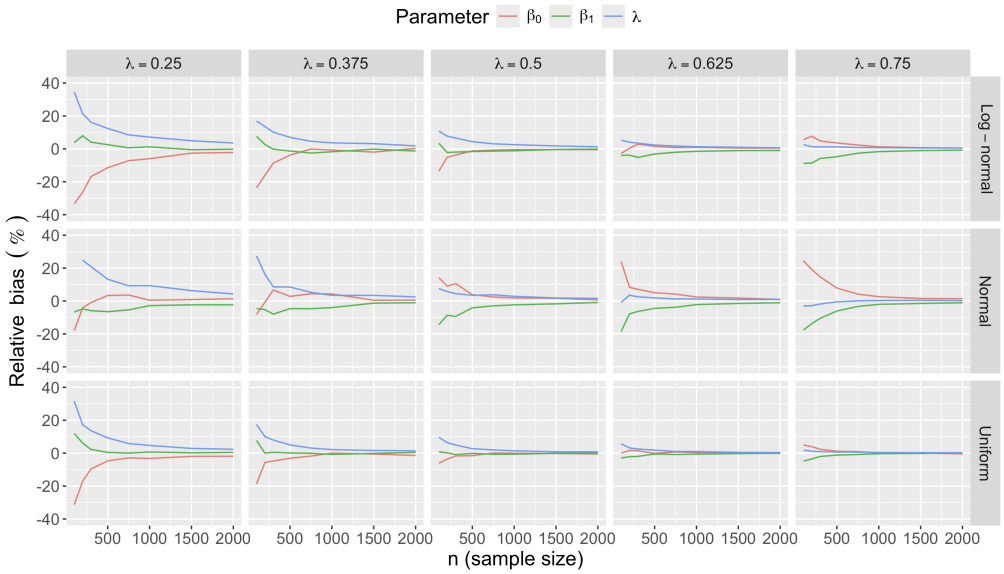


Figure A4. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -1$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

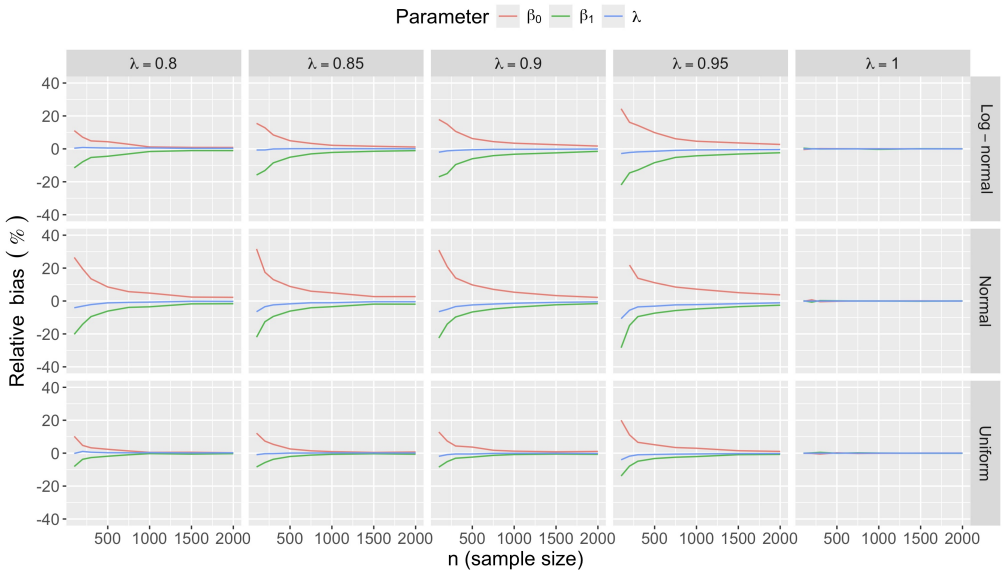


Figure A5. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -2$, and $\lambda \geq 0.8$.

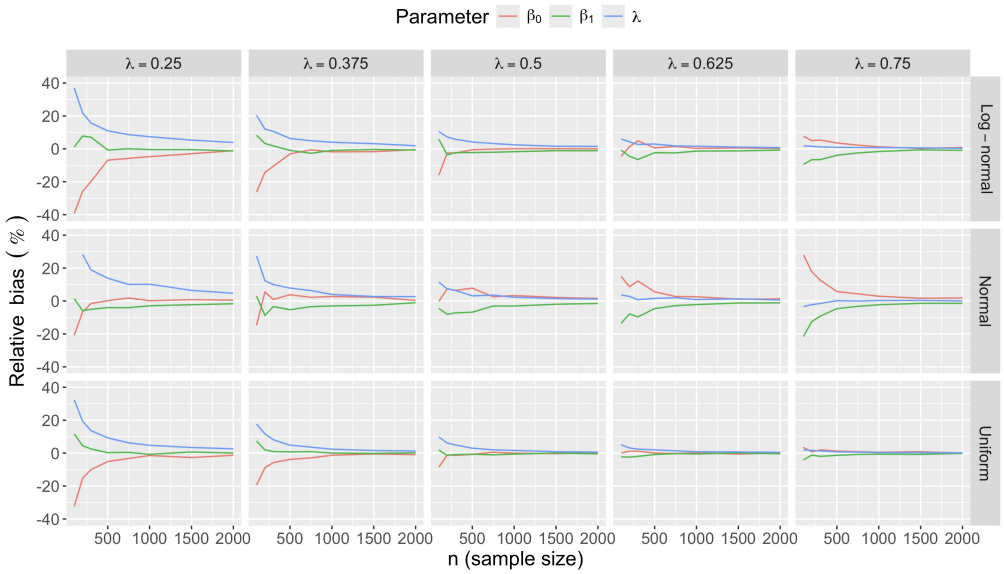


Figure A6. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -2$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

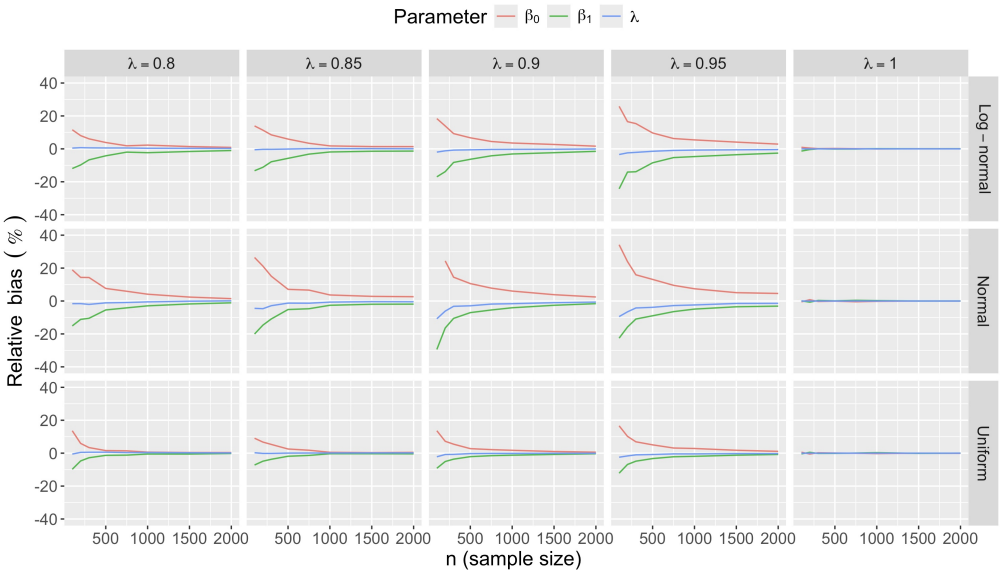


Figure A7. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -3$, and $\lambda \geq 0.8$.

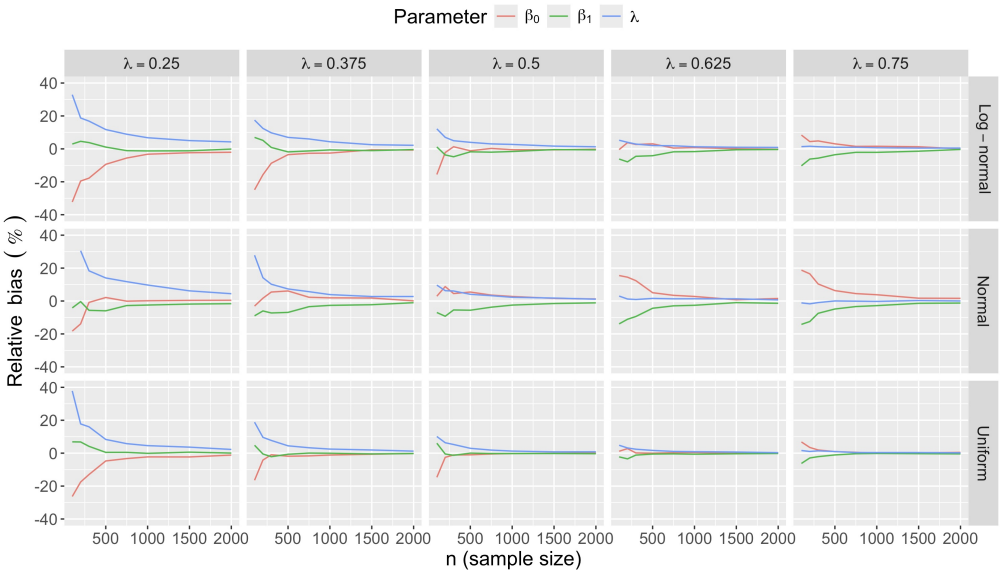


Figure A8. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -3$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

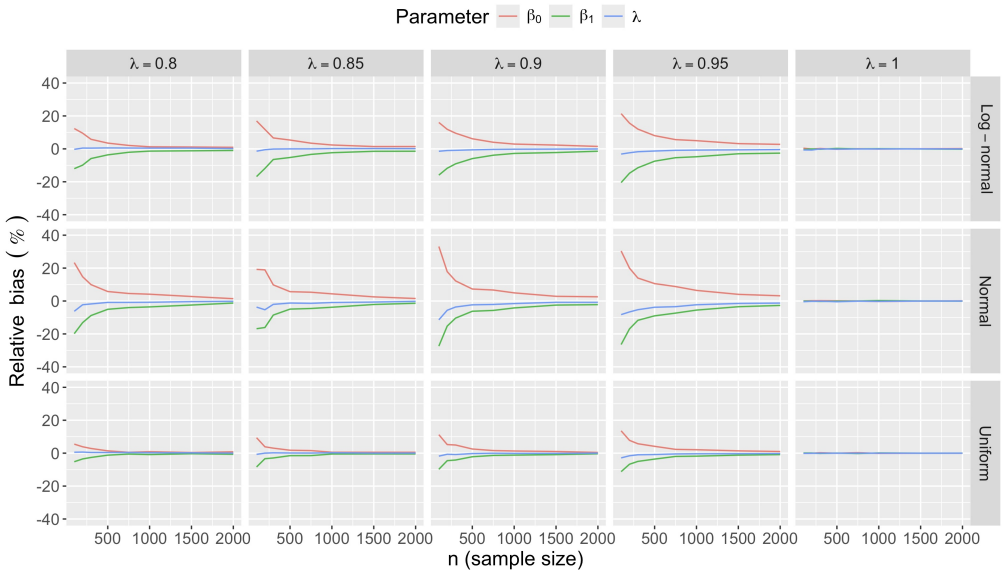


Figure A9. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -1$, and $\lambda \geq 0.8$.

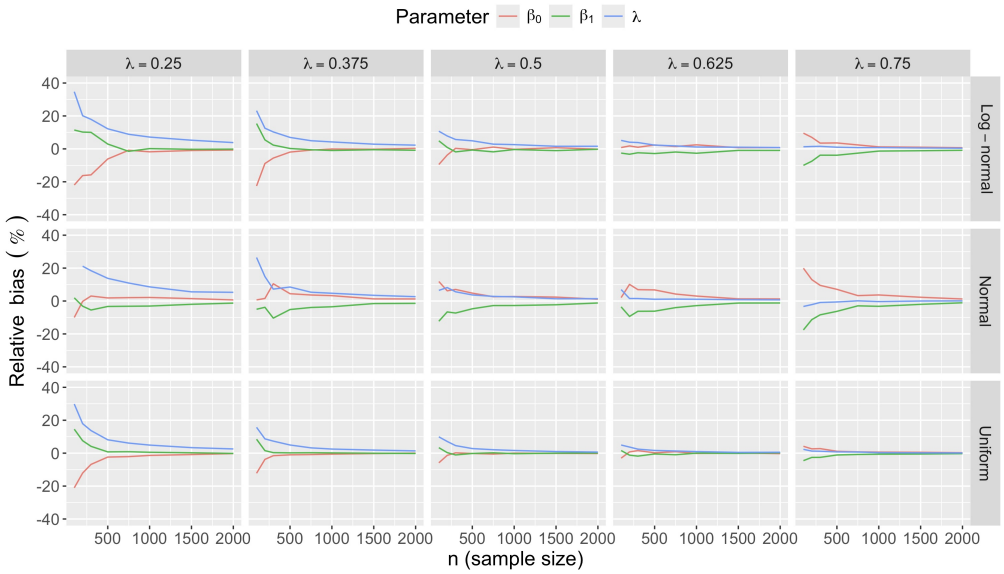


Figure A10. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -1$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

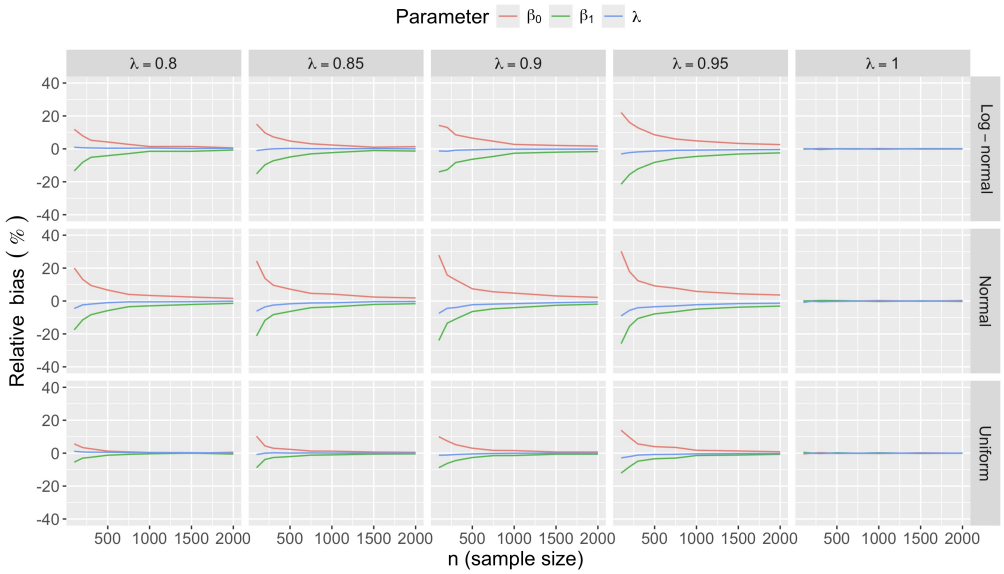


Figure A11. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -2$, and $\lambda \geq 0.8$.

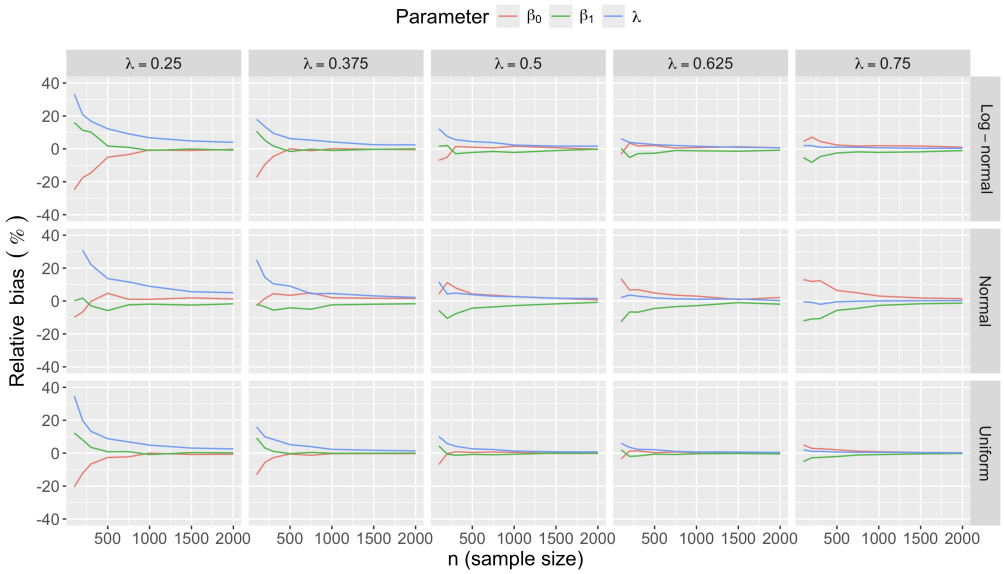


Figure A12. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -2$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

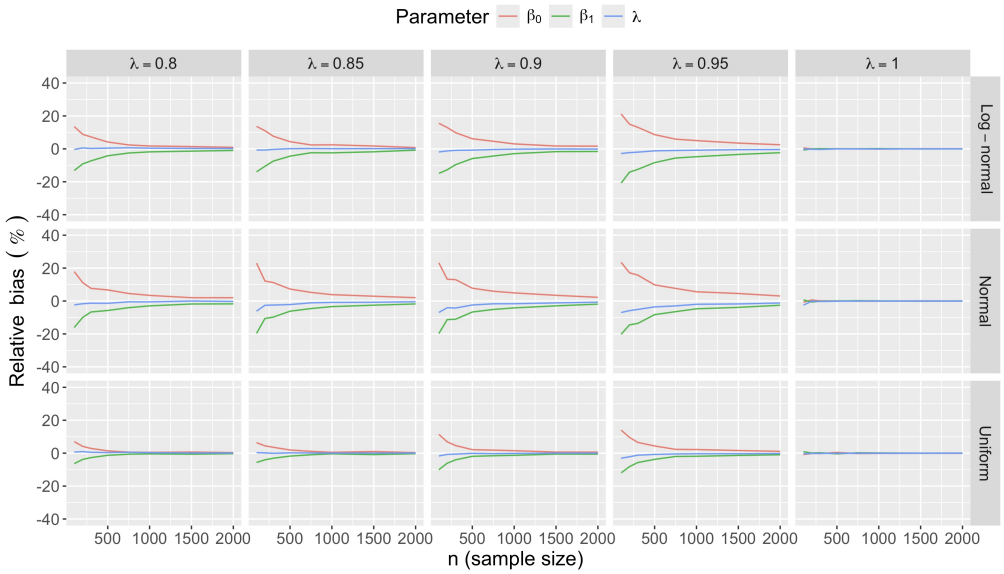


Figure A13. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -3$, and $\lambda \geq 0.8$.

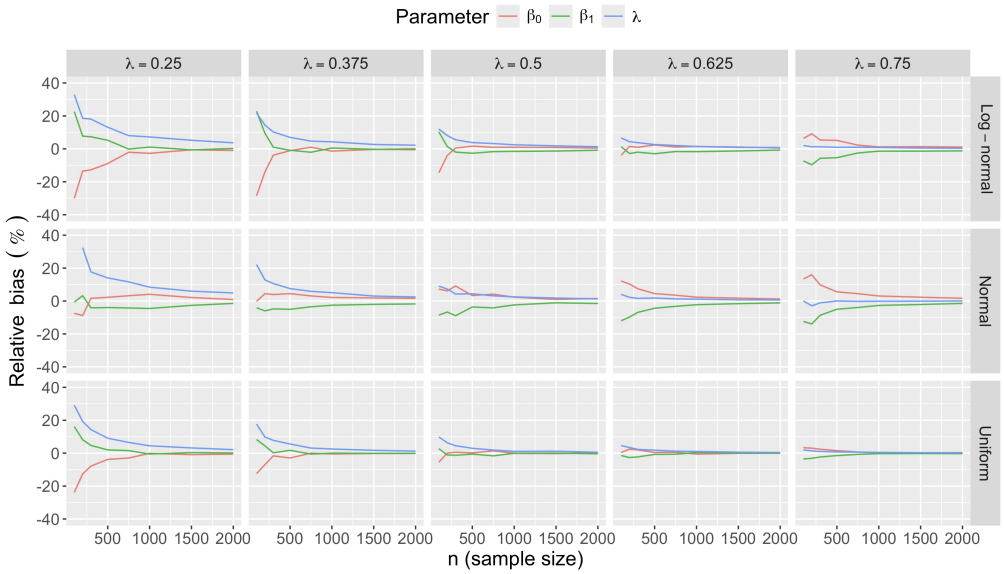


Figure A14. Relative bias of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -3$, and $\lambda \leq 0.75$.

Appendix C.2.2. Accuracy

Figures A15–A26 present the relative root mean square error (*rrmse*) in the estimates $\hat{\lambda}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ for model parameters $\lambda \in [0.25, 1]$, $\beta_0 \in \{3, 8\}$ and $\beta_1 \in \{-3, -2, -1\}$ and sample size $n \in [100, 2000]$. Figure A27 and Figure A28 shows the *rrmse* in $\hat{\beta}_1$ grouped for all parameter settings. The *rrmse* generally approaches zero as n increases.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

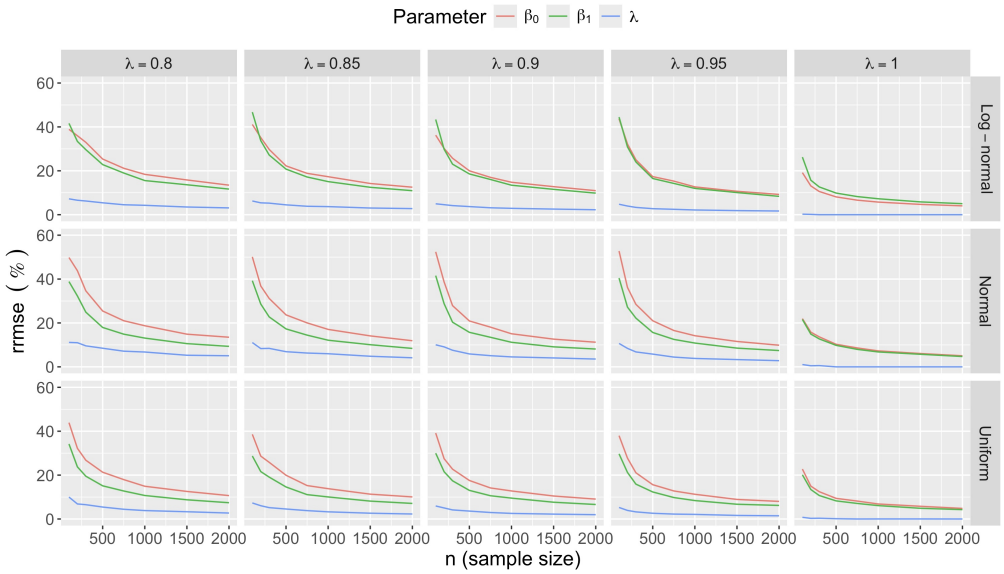


Figure A15. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -1$, and $\lambda \geq 0.8$.

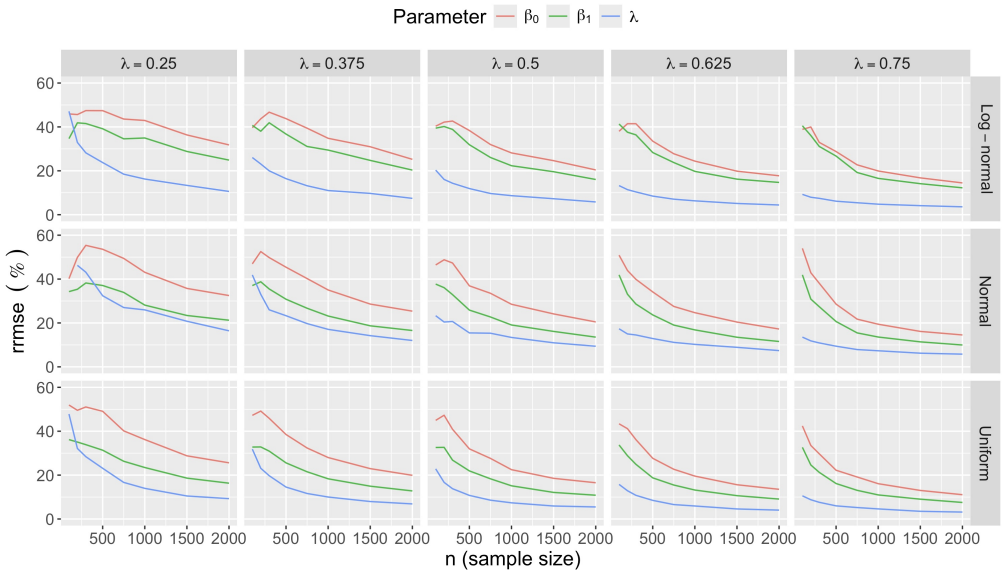


Figure A16. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -1$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

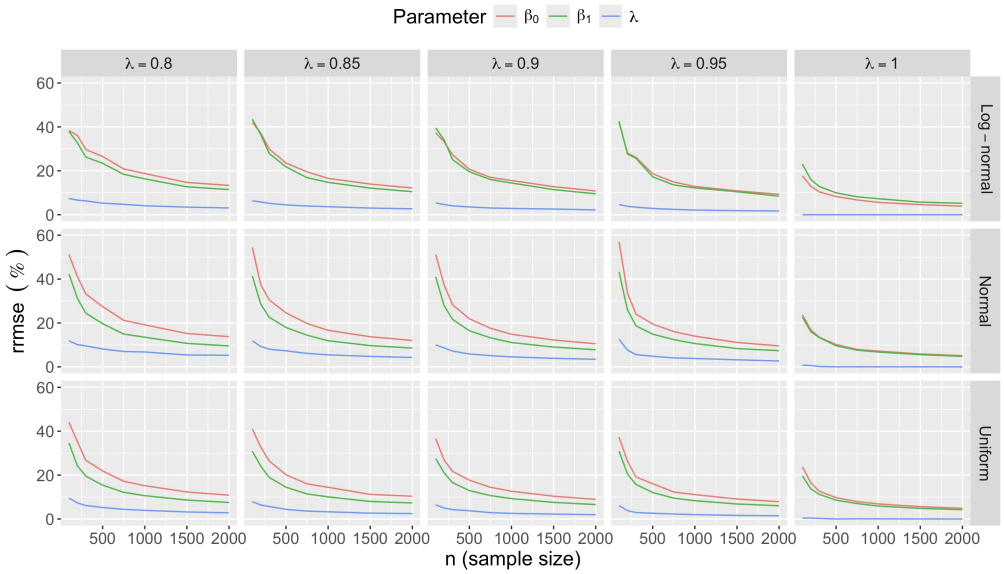


Figure A17. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -2$, and $\lambda \geq 0.8$.

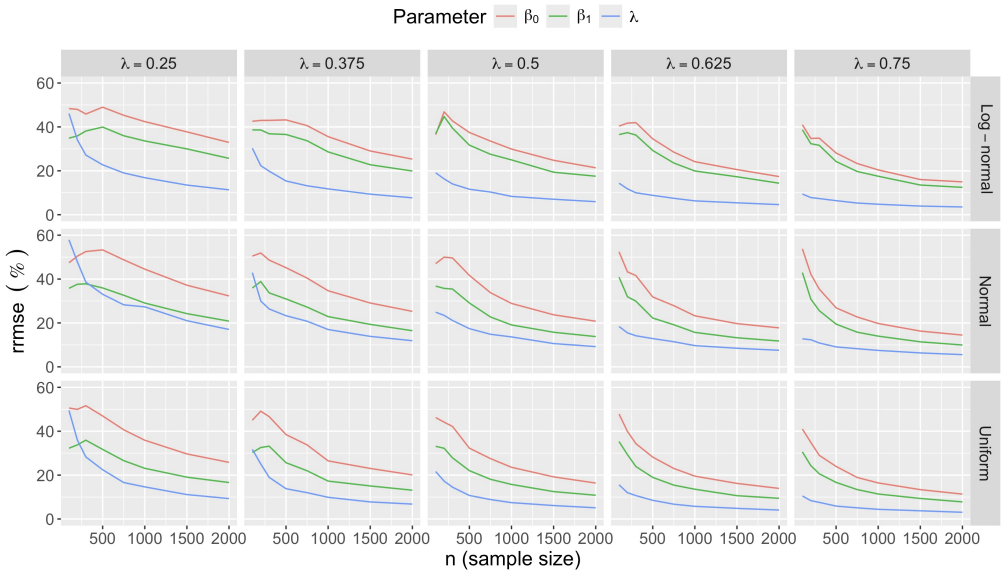


Figure A18. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -2$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

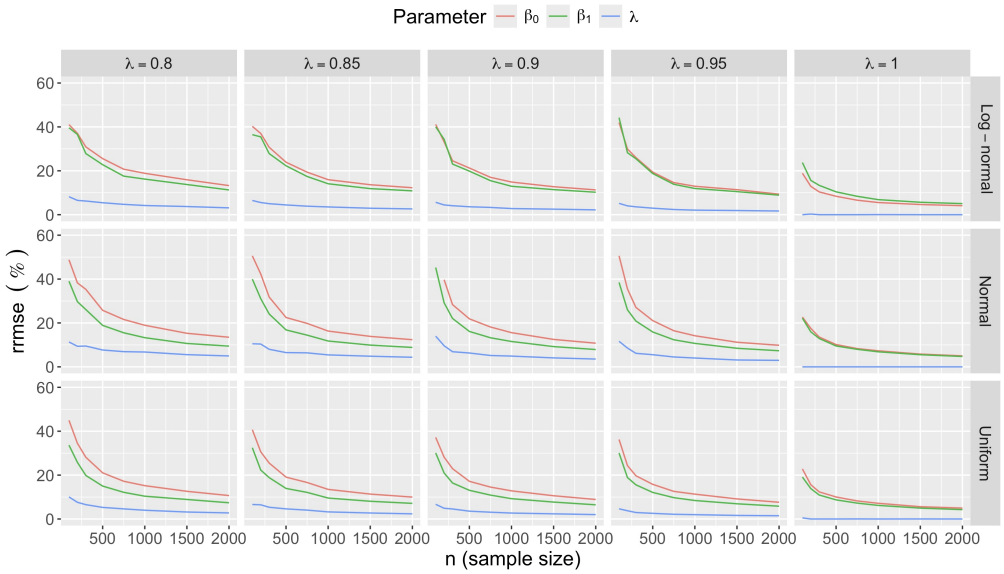


Figure A19. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -3$, and $\lambda \geq 0.8$.

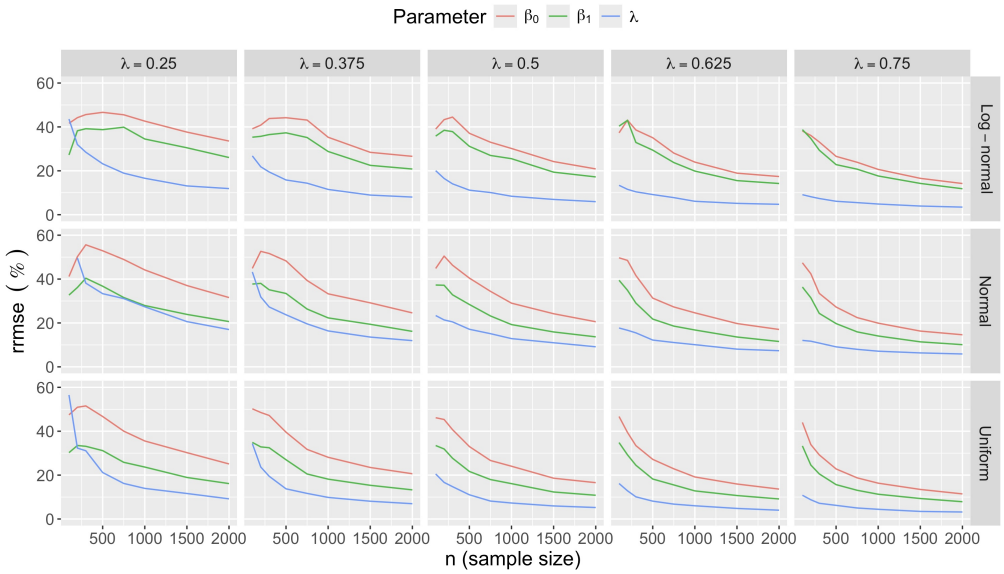


Figure A20. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -3$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

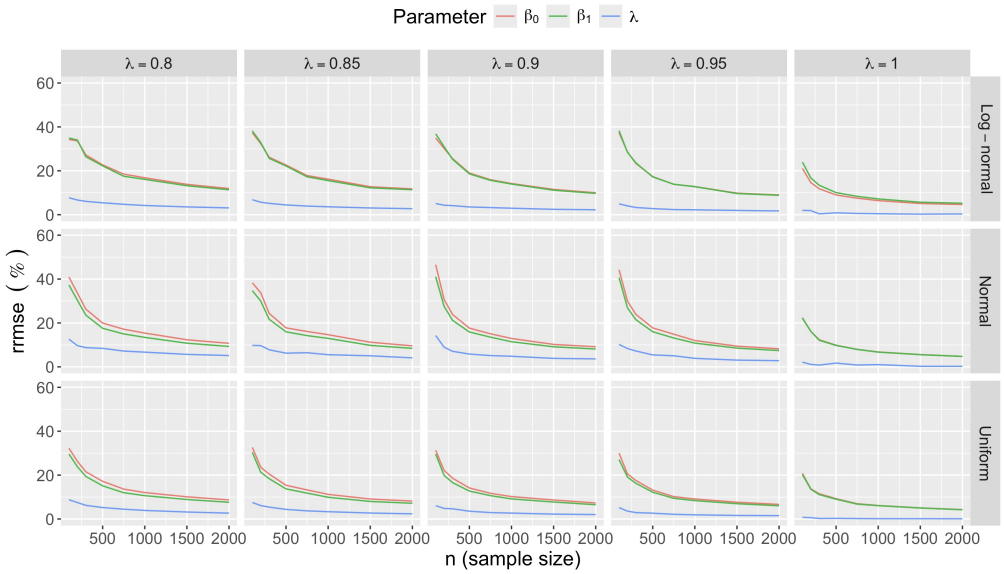


Figure A21. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8, \beta_1 = -1$, and $\lambda \geq 0.8$.

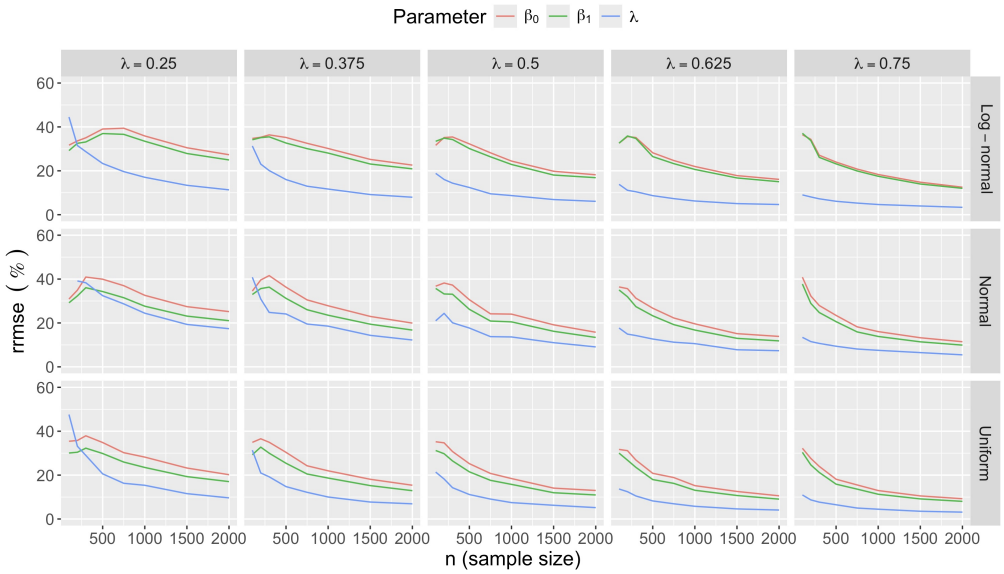


Figure A22. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8, \beta_1 = -1$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

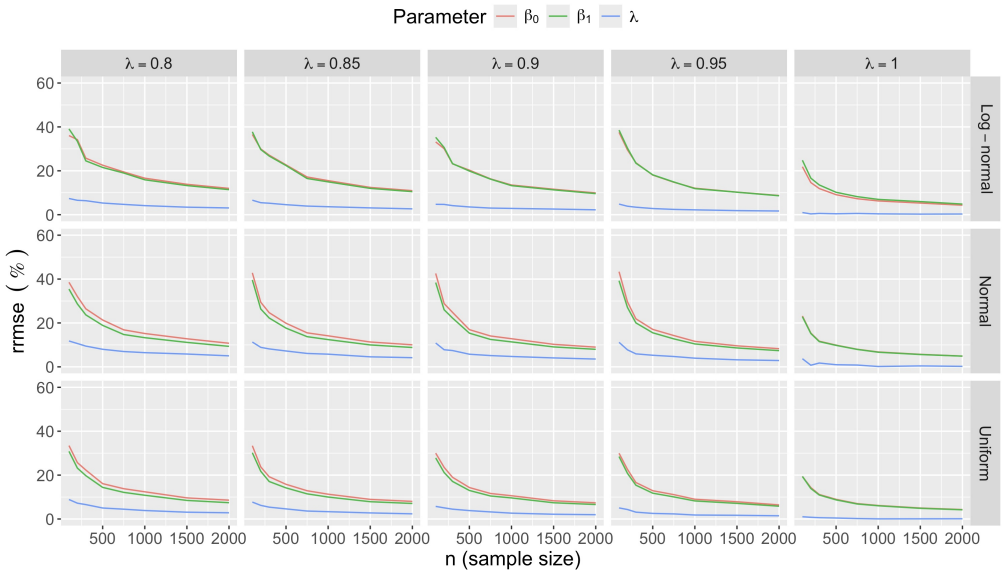


Figure A23. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -2$, and $\lambda \geq 0.8$.

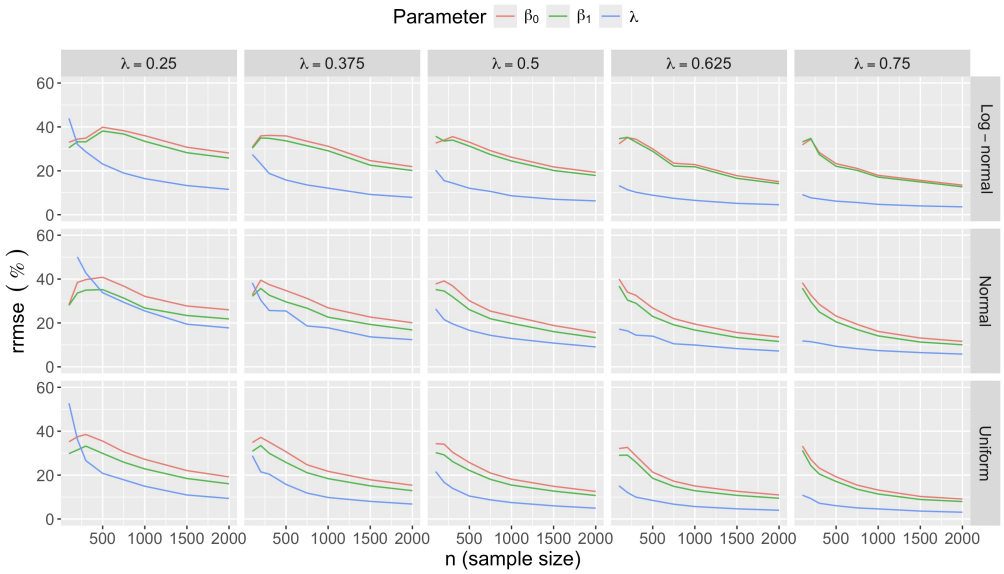


Figure A24. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -2$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

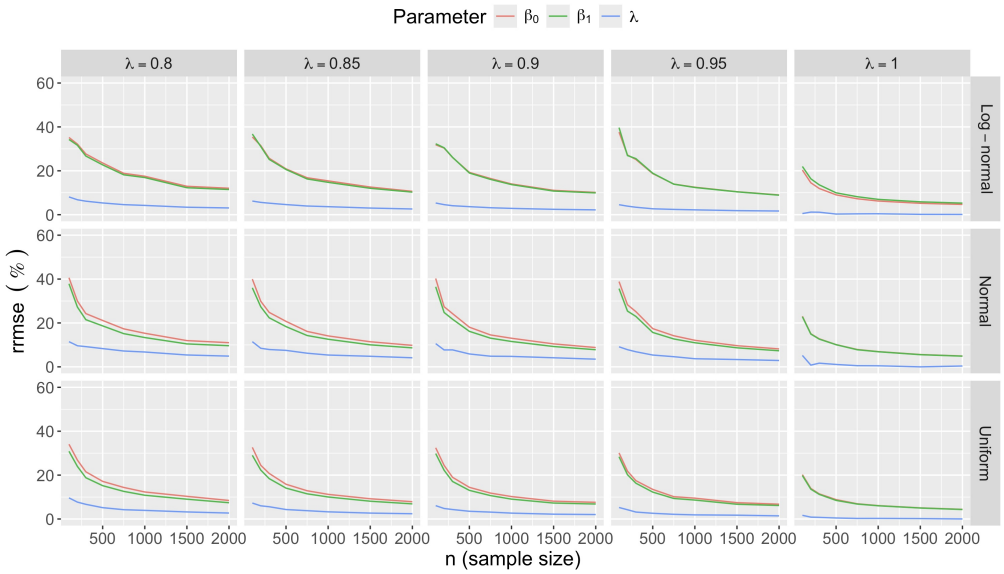


Figure A25. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -3$, and $\lambda \geq 0.8$.

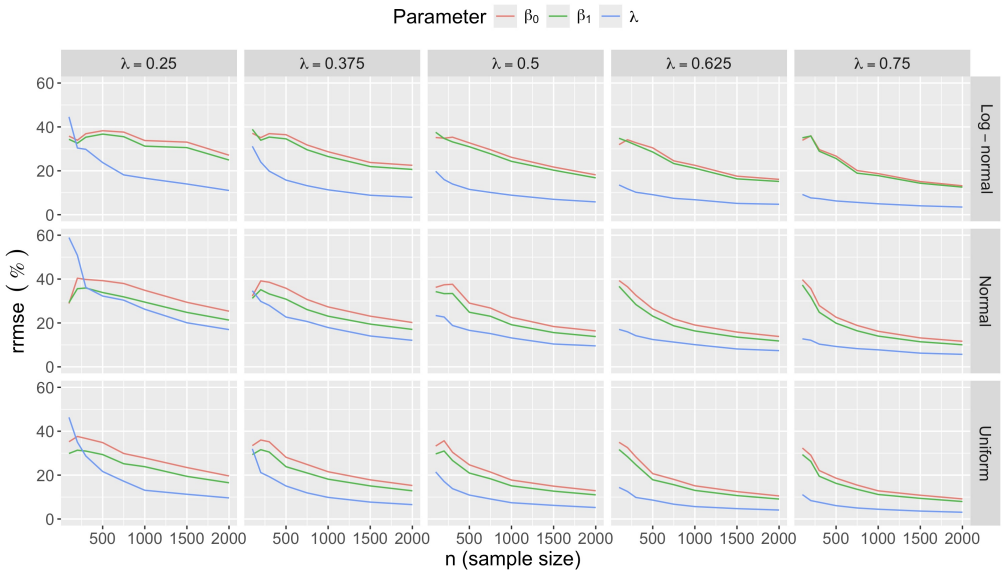


Figure A26. Relative root mean square error (*rmse*) of Penalized Maximum Likelihood estimates of multistage binomial model parameters λ, β_0 and β_1 under the MSB model data generation process in Equation (A22) with $\beta_0 = 8$, $\beta_1 = -3$, and $\lambda \leq 0.75$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

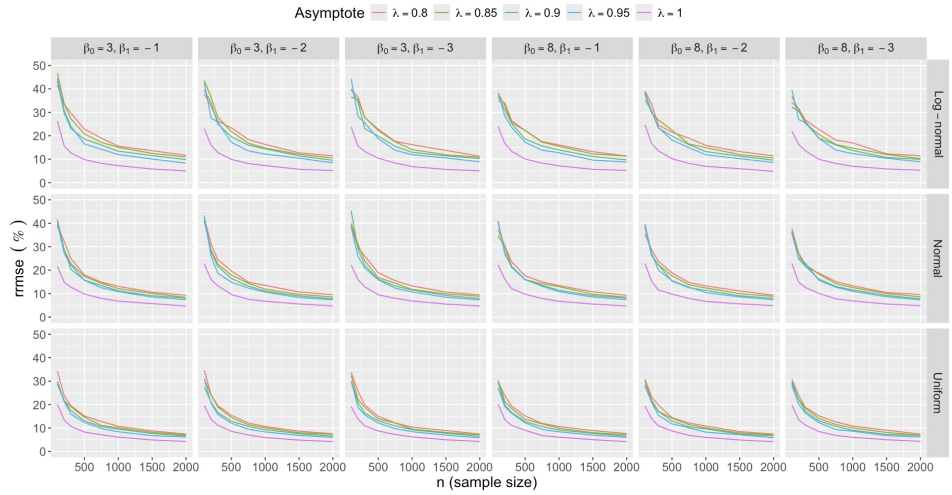


Figure A27. Relative root mean square error (*rrmse*) of Penalized Maximum Likelihood estimate of the slope β_1 under the MSB data generation process in Equation (A22) with $q = 1$ (one predictor) for $\lambda \geq 0.8$. The figure columns show the investigated pairs of true β_0 and β_1 values, and rows correspond to investigated distributions of predictor.

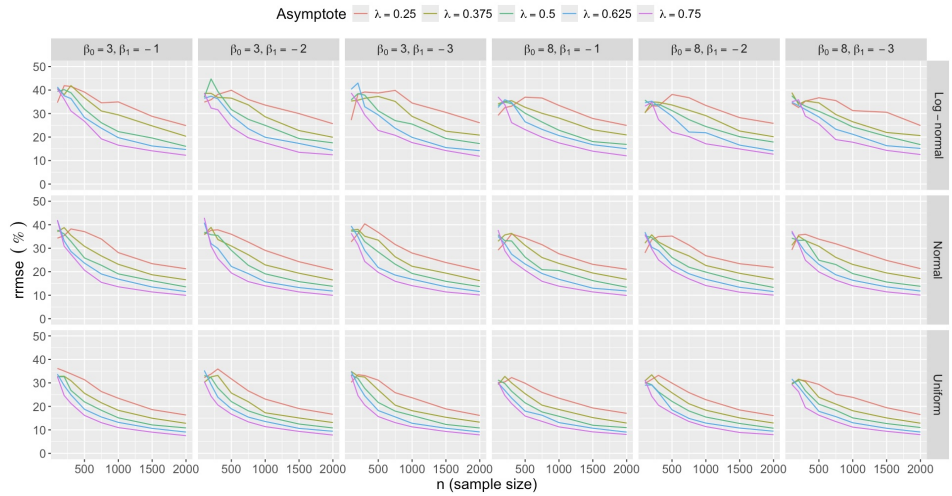


Figure A28. Relative root mean square error (*rrmse*) of Penalized Maximum Likelihood estimate of the slope β_1 under the MSB data generation process in Equation (A22) with $q = 1$ (one predictor) for $\lambda \leq 0.75$. The figure columns show the investigated pairs of true β_0 and β_1 values, and rows correspond to investigated distributions of predictor.

Appendix C.2.3. Sample Size Determination

In complement to Figure 4, Figures A29 and A30 present the sample size n required to achieve a specified relative root mean square error for the PML estimate $\hat{\beta}_1$ of the regression slope β_1 under the MSB data generation process in Equation (A22) with $\beta_0 = 3$, $\beta_1 = -2$ for normal and log-normal predictor distributions, respectively.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

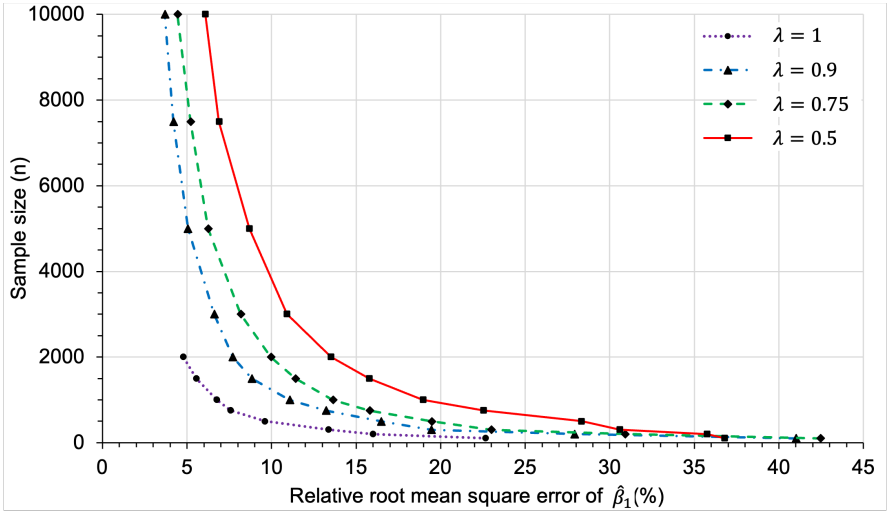


Figure A29. Sample size n required to achieve a specified relative root mean square error for the Penalized Maximum Likelihood estimate $\hat{\beta}_1$ of the regression slope β_1 under the MSB data generation process in Equation (A22) with a normal predictor distribution, $\beta_0 = 3$, $\beta_1 = -2$ and selected λ values.

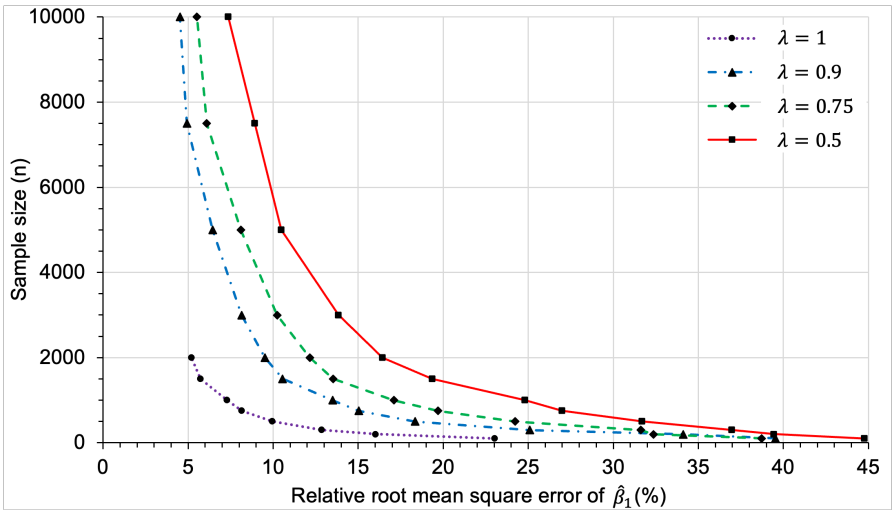


Figure A30. Sample size n required to achieve a specified relative root mean square error for the Penalized Maximum Likelihood estimate $\hat{\beta}_1$ of the regression slope β_1 under the MSB data generation process in Equation (A22) with a log-normal predictor distribution, $\beta_0 = 3$, $\beta_1 = -2$ and selected λ values.

References

1. Harris, J.K. Primer on binary logistic regression. *Family medicine and community health* **2021**, *9*, e001290.
2. Zabor, E.C.; Reddy, C.A.; Tendulkar, R.D.; Patil, S. Logistic regression in clinical studies. *International Journal of Radiation Oncology* Biology* Physics* **2022**, *112*, 271–277.
3. Halvorson, M.A.; McCabe, C.J.; Kim, D.S.; Cao, X.; King, K.M. Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. *Psychology of Addictive Behaviors* **2022**, *36*, 284.
4. Nelder, J.A.; Wedderburn, R.W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **1972**, *135*, 370–384.
5. Webster, A.J. Multi-stage models for the failure of complex systems, cascading disasters, and the onset of disease. *PloS one* **2019**, *14*, e0216422.
6. Fröhlich, F.; Theis, F.J.; Rädler, J.O.; Hasenauer, J. Parameter estimation for dynamical systems with discrete events and logical operations. *Bioinformatics* **2017**, *33*, 1049–1056.
7. Zhang, T.; Rohlfs, R.; Schwartz, R. Implementation of a discrete event simulator for biological self-assembly systems. In Proceedings of the Proceedings of the Winter Simulation Conference, 2005. IEEE, 2005, pp. 9–pp.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

8. Manhart, M.; Morozov, A.V. Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proceedings of the National Academy of Sciences* **2015**, *112*, 1797–1802.
9. Scarborough, P.; Nnoaham, K.E.; Clarke, D.; Capewell, S.; Rayner, M. Modelling the impact of a healthy diet on cardiovascular disease and cancer mortality. *J Epidemiol Community Health* **2012**, *66*, 420–426.
10. Walter, S.; Holford, T. Additive, multiplicative, and other models for disease risks. *American Journal of Epidemiology* **1978**, *108*, 341–346.
11. Zhang, X.; Fröhlich, H.; Grigoriev, D.; Vakulenko, S.; Zimmermann, J.; Weber, A.G. A simple 3-parameter model for cancer incidences. *Scientific Reports* **2018**, *8*, 3388.
12. Wu, S.; Powers, S.; Zhu, W.; Hannun, Y.A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **2016**, *529*, 43–47.
13. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. *International Journal of Epidemiology* **2016**, *45*, 621–637.
14. Wacholder, S.; Han, S.S.; Weinberg, C.R. Inference from a multiplicative model of joint genetic effects for ovarian cancer risk. *Journal of the National Cancer Institute* **2011**, *103*, 82–83.
15. Calabrese, P.; Shibata, D. A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates. *BMC cancer* **2010**, *10*, 1–12.
16. Siemiatycki, J.; Thomas, D.C. Biological models and statistical interactions: an example from multistage carcinogenesis. *International journal of epidemiology* **1981**, *10*, 383–387.
17. Armitage, P. Multistage models of carcinogenesis. *Environmental health perspectives* **1985**, *63*, 195–201.
18. Loomba, R.; Yang, H.I.; Su, J.; Brenner, D.; Iloeje, U.; Chen, C.J. Obesity and alcohol synergize to increase the risk of incident hepatocellular carcinoma in men. *Clinical Gastroenterology and Hepatology* **2010**, *8*, 891–898.
19. Ishak, K.J.; Kreif, N.; Benedict, A.; Muszbek, N. Overview of parametric survival analysis for health-economic applications. *Pharmacoeconomics* **2013**, *31*, 663–675.
20. Chang, Y.T.; Wu, J.L.; Hsu, C.C.; Wang, J.D.; Sung, J.M. Diabetes and end-stage renal disease synergistically contribute to increased incidence of cardiovascular events: a nationwide follow-up study during 1998–2009. *Diabetes care* **2014**, *37*, 277–285.
21. Ramírez-Aldana, R.; Gomez-Verjan, J.C.; Bello-Chavolla, O.Y.; García-Peña, C. Spatial epidemiological study of the distribution, clustering, and risk factors associated with early COVID-19 mortality in Mexico. *PloS one* **2021**, *16*, e0254884.
22. Lee, A.; Mavaddat, N.; Wilcox, A.N.; Cunningham, A.P.; Carver, T.; Hartley, S.; Babb de Villiers, C.; Izquierdo, A.; Simard, J.; Schmidt, M.K.; et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine* **2019**, *21*, 1708–1718.
23. Di Angelantonio, E.; Kaptoge, S.; Wormser, D.; Willeit, P.; Butterworth, A.S.; Bansal, N.; O'Keeffe, L.M.; Gao, P.; Wood, A.M.; Burgess, S.; et al. Association of cardiometabolic multimorbidity with mortality. *Jama* **2015**, *314*, 52–60.
24. Ott, J.; Ullrich, A.; Mascarenhas, M.; Stevens, G. Global cancer incidence and mortality caused by behavior and infection. *Journal of public health* **2011**, *33*, 223–233.
25. Lee, P. Relation between exposure to asbestos and smoking jointly and the risk of lung cancer. *Occupational and environmental medicine* **2001**, *58*, 145–153.
26. Katz, K.A. The (relative) risks of using odds ratios. *Archives of dermatology* **2006**, *142*, 761–764.
27. Greenland, S. Noncollapsibility, confounding, and sparse-data bias. Part 1: The oddities of odds. *Journal of clinical epidemiology* **2021**, *138*, 178–181.
28. Greenland, S. Noncollapsibility, confounding, and sparse-data bias. Part 2: What should researchers make of persistent controversies about the odds ratio? *Journal of clinical epidemiology* **2021**, *139*, 264–268.
29. Burgess, S. Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. *Communications in Statistics-Theory and Methods* **2017**, *46*, 786–804.
30. Janani, L.; Mansournia, M.A.; Nourijeylani, K.; Mahmoodi, M.; Mohammad, K. Statistical issues in estimation of adjusted risk ratio in prospective studies. *Archives of Iranian medicine* **2015**, *18*, 0–0.
31. Correia, K.; Williams, P.L. Estimating the relative excess risk due to interaction in clustered-data settings. *American Journal of Epidemiology* **2018**, *187*, 2470–2480.
32. Marschner, I.C. Relative risk regression for binary outcomes: methods and recommendations. *Australian & New Zealand Journal of Statistics* **2015**, *57*, 437–462.
33. Dwivedi, A.K.; Mallawaarachchi, I.; Lee, S.; Tarwater, P. Methods for estimating relative risk in studies of common binary outcomes. *Journal of applied statistics* **2014**, *41*, 484–500.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

34. Chu, H.; Nie, L.; Chen, Y.; Huang, Y.; Sun, W. Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. *Statistical methods in medical research* **2012**, *21*, 621–633.
35. Hall, D.B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **2000**, *56*, 1030–1039.
36. Richardson, T.S.; Robins, J.M.; Wang, L. On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association* **2017**, *112*, 1121–1130.
37. Yin, J.; Markes, S.; Richardson, T.S.; Wang, L. Multiplicative effect modelling: the general case. *Biometrika* **2022**, *109*, 559–566.
38. Doi, S.A.; Furuya-Kanamori, L.; Xu, C.; Lin, L.; Chivese, T.; Thalib, L. Controversy and debate: questionable utility of the relative risk in clinical research: paper 1: a call for change to practice. *Journal of clinical epidemiology* **2022**, *142*, 271–279.
39. Xiao, M.; Chen, Y.; Cole, S.R.; MacLehose, R.F.; Richardson, D.B.; Chu, H. Controversy and Debate: Questionable utility of the relative risk in clinical research: Paper 2: Is the Odds Ratio "portable" in meta-analysis? Time to consider bivariate generalized linear mixed model. *Journal of Clinical Epidemiology* **2022**, *142*, 280–287.
40. Xiao, M.; Chu, H.; Cole, S.R.; Chen, Y.; MacLehose, R.F.; Richardson, D.B.; Greenland, S. Controversy and Debate: Questionable utility of the relative risk in clinical research: Paper 4: Odds Ratios are far from "portable"—A call to use realistic models for effect variation in meta-analysis. *Journal of Clinical Epidemiology* **2022**, *142*, 294–304.
41. Suhail A, D.; Furuya-Kanamori, L.; Xu, C.; Lin, L.; Chivese, T.; Thalib, L. Questionable utility of the relative risk in clinical research: A call for change to practice. *J Clin Epidemiol* **2020**.
42. Mood, C. Logistic regression: Uncovering unobserved heterogeneity, 2017. 341160.
43. Mood, C. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review* **2010**, *26*, 67–82.
44. Cramer, J.S. Omitted variables and misspecified disturbances in the logit model. Technical report, Tinbergen Institute Discussion Paper, 2005.
45. Yatchew, A.; Griliches, Z. Specification error in probit models. *The Review of Economics and Statistics* **1985**, pp. 134–139.
46. Lee, L.F. Specification error in multinomial logit models: Analysis of the omitted variable bias. *Journal of Econometrics* **1982**, *20*, 197–209.
47. Buis, M.L. Logistic regression: When can we do what we think we can do, 2017.
48. Brzoska, P.; Sauzet, O.; Breckenkamp, J. Unobserved heterogeneity and the comparison of coefficients across nested logistic regression models: how to avoid comparing apples and oranges. *International Journal of Public Health* **2017**, *62*, 517–520.
49. Cologne, J.; Furukawa, K.; Grant, E.J.; Abbott, R.D. Effects of omitting non-confounding predictors from general relative-risk models for binary outcomes. *Journal of Epidemiology* **2019**, *29*, 116–122.
50. Schuster, N.A.; Twisk, J.W.; Ter Riet, G.; Heymans, M.W.; Rijnhart, J.J. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC medical research methodology* **2021**, *21*, 1–9.
51. Berkson, J. Are there two regressions? *Journal of the american statistical association* **1950**, *45*, 164–180.
52. King, G.; Zeng, L. Logistic regression in rare events data. *Political analysis* **2001**, *9*, 137–163.
53. Huang, F.L. Alternatives to logistic regression models in experimental studies. *The Journal of Experimental Education* **2022**, *90*, 213–228.
54. Greenland, S.; Mansournia, M.A. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in medicine* **2015**, *34*, 3133–3143.
55. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **1993**, *80*, 27–38.
56. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1984**, *71*, 1–10.
57. Heinze, G.; Schemper, M. A solution to the problem of separation in logistic regression. *Statistics in medicine* **2002**, *21*, 2409–2419.
58. Zeng, G.; Zeng, E. On the relationship between multicollinearity and separation in logistic regression. *Communications in Statistics-Simulation and Computation* **2021**, *50*, 1989–1997.
59. Mansournia, M.A.; Geroldinger, A.; Greenland, S.; Heinze, G. Separation in logistic regression: causes, consequences, and control. *American journal of epidemiology* **2018**, *187*, 864–870.
60. Rainey, C. Dealing with separation in logistic regression models. *Political Analysis* **2016**, *24*, 339–355.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

61. Botes, M.; Fletcher, L. Comparing logistic regression methods for a sparse data set when complete separation is present. In Proceedings of the Annual Proceedings of the South African Statistical Association Conference. South African Statistical Association (SASA), 2014, number con-1 in 2014, pp. 1–8.
62. Hauck Jr, W.W.; Donner, A. Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association* **1977**, *72*, 851–853.
63. Diop, A.; Diop, A.; Dupuy, J.F. Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electronic Journal of Statistics* **2011**, *5*, 460–483.
64. Diop, A.; Diop, A.; Dupuy, J.F. Simulation-based inference in a zero-inflated Bernoulli regression model. *Communications in Statistics-Simulation and Computation* **2016**, *45*, 3597–3614.
65. Diallo, A.O.; Diop, A.; Dupuy, J.F. Asymptotic properties of the maximum-likelihood estimator in zero-inflated binomial regression. *Communications in Statistics-Theory and Methods* **2017**, *46*, 9930–9948.
66. Pho, K.H. Goodness of fit test for a zero-inflated Bernoulli regression model. *Communications in Statistics-Simulation and Computation* **2024**, *53*, 756–771.
67. Diop, A.; Ba, D.B.; Lo, F. Asymptotic properties in the Probit-Zero-inflated Binomial regression model. *arXiv preprint arXiv:2105.00483* **2021**.
68. Garel, M.; Solberg, E.J.; Sæther, B.E.; Grøtan, V.; Tufto, J.; Heim, M. Age, size, and spatiotemporal variation in ovulation patterns of a seasonal breeder, the Norwegian moose (Alces alces). *The American Naturalist* **2009**, *173*, 89–104.
69. Longford, N. Analysis of Covariance. In *International Encyclopedia of Education (Third Edition)*, Third Edition ed.; Peterson, P.; Baker, E.; McGaw, B., Eds.; Elsevier: Oxford, 2010; pp. 18–24. https://doi.org/https://doi.org/10.1016/B978-0-08-044894-7.01304-X.
70. Pan, J.; Thompson, R. Gauss-Hermite quadrature approximation for estimation in generalised linear mixed models. *Computational statistics* **2003**, *18*, 57–78.
71. Rabe-Hesketh, S.; Skrondal, A.; Pickles, A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* **2002**, *2*, 1–21.
72. Asheri, H.; Hosseini, R.; Araabi, B.N. A new EM algorithm for flexibly tied GMMs with large number of components. *Pattern recognition* **2021**, *114*, 107836.
73. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)* **1977**, *39*, 1–22.
74. Wang, Z.; Louis, T.A. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* **2003**, *90*, 765–775.
75. Azzalini, A.; Valle, A.D. The multivariate skew-normal distribution. *Biometrika* **1996**, *83*, 715–726.
76. Ghosh, P. Corrected Probability Mass Function of the Binomial Distribution. *Resonance* **2021**, *26*, 1721–1724.
77. Shklyar, S. Identifiability of logistic regression with homoscedastic error: Berkson model. *Modern Stochastics: Theory and Applications* **2015**, *2*, 131–146.
78. Küchenhoff, H. The identification of logistic regression models with errors in the variables. *Statistical Papers* **1995**, *36*, 41–47.
79. Al-Shaaibi, M.; Wesonga, R. Bias reduction in the logistic model parameters with the LogF (1, 1) penalty under MAR assumption. *Frontiers in Applied Mathematics and Statistics* **2022**, *8*, 1052752.
80. Heinze, G.; Puhr, R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in medicine* **2010**, *29*, 770–777.
81. Heinze, G.; Ploner, M. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer methods and programs in biomedicine* **2003**, *71*, 181–187.
82. Jeffreys, H. *The theory of probability*; OuP Oxford, 1998.
83. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* **1922**, *222*, 309–368.
84. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025.
85. Susko, E. Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika* **2013**, *100*, 1019–1023.
86. Shapiro, A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **1985**, *72*, 133–144.
87. Molenberghs, G.; Verbeke, G. Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician* **2007**, *61*, 22–27.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2025 doi:10.20944/preprints202509.2376.v1

88. Mitchell, J.D.; Allman, E.S.; Rhodes, J.A. Hypothesis testing near singularities and boundaries. *Electronic journal of statistics* **2019**, *13*, 2150.
89. Baey, C.; Cournède, P.H.; Kuhn, E. Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics & Data Analysis* **2019**, *135*, 107–122.
90. Xu, C.; Bull, S.B. Penalized maximum likelihood inference under the mixture cure model in sparse data. *Statistics in Medicine* **2023**, *42*, 2134–2161.
91. Drton, M.; Williams, B. Quantifying the failure of bootstrap likelihood ratio tests. *Biometrika* **2011**, *98*, 919–934.
92. Cameron, A.C.; Windmeijer, F.A. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics* **1997**, *77*, 329–342.
93. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* **2019**, *38*, 2074–2102.
94. Cribari-Neto, F.; Zeileis, A. Beta Regression in R. *Journal of Statistical Software* **2010**, *34*, 1–24. https://doi.org/10.18637/jss.v034.i02.
95. Gueorguieva, R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical modelling* **2001**, *1*, 177–193.
96. Bonat, W.H.; Jørgensen, B. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society Series C: Applied Statistics* **2016**, *65*, 649–675.
97. Kosmidis, I.; Firth, D. Bias reduction in exponential family nonlinear models. *Biometrika* **2009**, *96*, 793–804.
98. Kosmidis, I.; Firth, D. A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics* **2010**, *4*, 1097–1112.
99. Blizzard, L.; Hosmer, W. Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal* **2006**, *48*, 5–22.
100. Bartlett, M.S. Properties of sufficiency and statistical tests. *Proceedings of the royal society of london. series a-mathematical and physical sciences* **1937**, *160*, 268–282.
101. Rao, C. Score test: historical review and recent developments. *Advances in ranking and selection, multiple comparisons, and reliability: methodology and applications* **2005**, pp. 3–20.
102. Tovissodé, C.F.; Diop, A.; Glèlè Kakaï, R. Inference in skew generalized t-link models for clustered binary outcome via a parameter-expanded EM algorithm. *Plos one* **2021**, *16*, e0249604.
103. Vrabel, R. A note on the matrix determinant lemma. *International Journal of Pure and Applied Mathematics* **2016**, *111*, 643–646.
104. Guyon, X. *Statistique et Econométrie : du modèle linéaire aux modèles non-linéaires*; Ellipses, 2001; p. 205. ISBN 2-7298-0842-6.
105. Efron, B.; Hinkley, D.V. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **1978**, *65*, 457–483.
106. Cao, X.; Spall, J.C. Relative performance of expected and observed Fisher information in covariance estimation for maximum likelihood estimates. In Proceedings of the 2012 American Control Conference (ACC). IEEE, 2012, pp. 1871–1876.
107. Yuan, X.; Spall, J.C. Confidence intervals with expected and observed fisher information in the scalar case. In Proceedings of the 2020 American Control Conference (ACC). IEEE, 2020, pp. 2599–2604.
108. Jiang, S.; Spall, J.C. Comparison between Expected and Observed Fisher Information in Interval Estimation. In Proceedings of the 2021 55th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2021, pp. 1–6.
109. Magnus, J.R.; Neudecker, H. Matrix differential calculus with applications to simple, Hadamard, and Kronecker products. *Journal of Mathematical Psychology* **1985**, *29*, 474–492.
110. Magnus, J.R.; Neudecker, H. The commutation matrix: some properties and applications. *The annals of statistics* **1979**, *7*, 381–394.
111. Hamid, H.A.; Wah, Y.B.; Xie, X.J.; Rahman, H.A.A. Assessing the effects of different types of covariates for binary logistic regression. In Proceedings of the AIP Conference Proceedings. American Institute of Physics, 2015, Vol. 1643, pp. 425–430.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.